Frame-Based Scene Understanding: Structured Representations for Introspective Perception in Autonomous Driving

Li Liu LLIU112@UCSC.EDU
Leilani H. Gilpin LGILPIN@UCSC.EDU
Computer Science and Engineering, UC Santa Cruz, Santa cruz, CA 95064 USA

Abstract

Autonomous driving systems excel at low-level perception but often lack structured, human-interpretable understanding of dynamic scenes, limiting transparency, robustness, and introspection. We present a cognitive framework for frame-based scene understanding, that transforms sensor-aligned observations into a hierarchical set of symbolic frames at the sample, object, and scene levels. Our pipeline constructs ego-referenced trajectories, applies rule-based behavior parsing, and produces natural language descriptors aligned with symbolic slot-filler structures. These representations support introspective capabilities such as expectation-driven anomaly detection, reasoning over uncertainty, and queryable explanations. We further integrate frames with large language models to probe symbolic-to-language reasoning tasks (summarization, intent inference, and counterfactuals) without raw sensor input. We describe implementation details, visualization tools, and application use cases, and outline an evaluation protocol combining qualitative case studies and task-based assessments. This work takes a step toward hybrid neuro-symbolic cognition for autonomy, enabling interpretable, reflective scene-understanding and human-aligned communication.

1. Introduction

Understanding traffic environments is a complex and high-stakes challenge in the domain of autonomous driving. It requires more than the detection of vehicles, pedestrians, and road infrastructure (Geiger et al., 2012; Cordts et al., 2016). Effective autonomy also requires the ability to interpret interactions between agents (Shi et al., 2022; Zhou et al., 2022; Seff et al., 2023), to predict their future behavior in dynamically evolving contexts (Chai et al., 2019; Salzmann et al., 2020), and to explain actions in terms of goals, constraints, and safety implications. Although recent advances in deep learning have produced high-performing models for object detection (Liu et al., 2023b; Li et al., 2024), multi-agent tracking (Zhang et al., 2022; Cao et al., 2023), and scene segmentation (Cheng et al., 2022; Tian et al., 2023; Li et al., 2023), these systems frequently operate as black boxes, and are not interpretable to humans (Lipton, 2018). They often lack semantic interpretability, provide no mechanisms for introspection, and fail to communicate their reasoning in human-understandable terms.

A key limitation of many current approaches for AVs are the reliance on end-to-end deep neural networks trained with bottom-up statistical learning. These systems build internal representations from raw sensor inputs such as RGB images and lidar point clouds (Du et al., 2023; Chen et al., 2024), constructing increasingly abstract features through stacked layers of nonlinear transformations. However, these learned features are sub-symbolic and distributed, making it difficult to inspect, generalize, or explain the model's behavior in unseen or ambiguous scenarios (Samek et al., 2017). In high-risk or safety-critical contexts, such opacity is a major liability (Amodei et al., 2016).

Because these models are optimized to exploit statistical correlations rather than structural or causal knowledge, they often fail in rare, out-of-distribution cases. They may behave unpredictably in edge situations, such as occluded intersections, erratic driver maneuvers, or abnormal pedestrian motion. Although techniques such as saliency visualization and attention maps (Selvaraju et al., 2017) have been proposed to make these systems more transparent, such tools only offer shallow insights. They do not explain why a model reached a particular conclusion, nor do they support counterfactual reasoning or structured queries.

As a result, there is growing recognition that perceptual competence alone is insufficient for robust autonomy. What is needed is a middle layer of abstraction that bridges low-level sensor data and high-level cognitive interpretation. This layer should represent context, causality, intent, and uncertainty in a structured and human-interpretable form. In this paper, we propose such a framework based on symbolic representations derived from **Frame Theory** (Minsky, 1974), a classic model of structured knowledge in cognitive science.

Frame Theory, as originally proposed by Marvin Minsky, conceptualizes stereotyped situations in terms of frames, which are structured collections of roles (or *slots*) and their expected values (or *fillers*). A frame may represent a common scenario such as "approaching an intersection" or "yielding to a pedestrian," and it can be instantiated with specific objects, agents, and temporal relationships. This structure supports default reasoning, anomaly detection, and introspection. Frames not only represent knowledge but also guide perception and action by activating expectations and identifying gaps in understanding.

We introduce a novel application of Frame Theory to the autonomous driving domain. Specifically, we present a pipeline that extracts symbolic, multi-level frame representations from the NuScenes dataset (Caesar et al., 2020), a large-scale benchmark featuring sensor-rich urban driving episodes. Our abstraction includes three levels of frames:

- **Sample-level frames**, which encode sensor-aligned, timestamped observations of objects and motion.
- **Object-level frames**, which aggregate temporal trajectories, classify agent behaviors, and track interactions with the ego vehicle.
- **Scene-level frames**, which summarize global context, traffic patterns, and scenario types across entire episodes.

These frames are designed to reflect cognitively meaningful units of perception and behavior. They organize raw sensor data into semantic representations that support introspective reasoning, symbolic querying, and explanation. We refer to this transformation as a form of *semantic compression*, where high-dimensional, noisy, and temporally distributed sensor signals are distilled into structured, human-interpretable knowledge templates.

Our central hypothesis is that such symbolic frame representations can serve as a "cognitive layer" between perception and decision-making modules in autonomous systems. They provide a substrate for symbolic reasoning, enable interpretability and debugging, and make it possible for large language models (LLMs) to engage with perceptual context in meaningful ways. Rather than conditioning LLMs on raw pixels or bounding boxes, we prompt them with symbolic descriptions of what occurred in a driving scene and explore whether they can simulate intent, infer causal structure, or generate counterfactual narratives.

This direction is inspired by recent findings that language models exhibit latent capabilities for physical reasoning, narrative inference, and theory-of-mind tasks (Huang et al., 2022; Kosinski, 2023; Bubeck et al., 2023). Our approach leverages this capacity by grounding symbolic inputs in a cognitively plausible representation, enabling reflective language-based reasoning over real-world driving events. In doing so, we seek to build AI agents that not only act on sensory data but also reason about their knowledge, limitations, and options, which is a hallmark of metacognition (Flavell, 1979; Schraw & Moshman, 1995).

From a broader cognitive systems perspective, our work connects with efforts to develop systems that can explain, imagine, and plan (Liang et al., 2016; Lake et al., 2017). In driving, hierarchical scene modeling and map-grounded forecasting further motivate structured representations (Wilson et al., 2023; Karnchanachari et al., 2024; Shi et al., 2025). We ask whether machines can imagine what happened in a traffic scenario given only a symbolic summary, and whether this imagination can inform safer behavior.

Contributions

To summarize, the key contributions of this work are as follows:

1. We formalize a hierarchical symbolic frame abstraction for modeling driving scenes, which encodes spatiotemporal semantics at multiple levels of granularity.

2. We develop a rule-based pipeline that extracts motion patterns and interaction descriptors from real-world lidar data, instantiating frames with natural language summaries.

3. We provide a protocol to evaluate whether large language models can reason over these symbolic frames, enabling scene interpretation, behavior prediction, and uncertainty reflection without raw visual input.

This work presents a preliminary yet functional implementation of a frame-based cognitive framework for autonomous scene understanding. The implemented components include: (i) a complete data-processing pipeline for frame construction at the sample, object, and scene levels; (ii) rule-based motion and interaction parsing from ego-aligned trajectories; and (iii) visualization modules for instance-level, map-based, and temporal analysis. The language-based reasoning and evaluation protocol using large language models are proposed as a next stage of investigation, to be developed and quantitatively validated in future work.

2. Related Work

2.1 Frame Theory and Semantic Representation

Frame Theory conceptualizes a stereotyped situation as a structured slot–filler scheme with default values that capture typical features and contextual expectations (Minsky, 1974). These structures allow an agent to map partial observations onto familiar contexts and then revise defaults when reality conflicts with expectations. Empirical research on perception and cognition shows that low-level sensory features are integrated into coherent object representations through attention (Treisman & Gelade, 1980), and that schematic frames support reasoning and transfer across different situations (Barsalou, 2012). In a driving scenario, a frame can encode roles such as the relative positions of vehicles, the intended behaviour of agents and environmental cues to represent situations like approaching an intersection or yielding to a pedestrian. By formalising such relations, frame-based representations provide a bridge between perception and higher-level reasoning that remains comprehensible to humans.

From a knowledge-representation perspective, it is not enough to record sensory data; a representation should support inference and problem solving. McCarthy and Hayes argued that AI systems need a general representation of the world that can be used to draw conclusions and plan actions (McCarthy & Hayes, 1981). Wilensky later emphasised that effective representations must expose the structural relations over which reasoning operates, rather than simply encoding raw data (Wilensky, 1987). Our hierarchical frame formulation reflects these principles: it explicitly encodes spatial and temporal relations among entities together with the goals or intentions of agents. This dual encoding echoes the multimodel approach of Chittaro and colleagues, who combine structural descriptions of a physical system with teleological information about its intended function to support reasoning and diagnosis (Chittaro et al., 1993). Recent work on abstraction schemes in computational game design formalises abstraction properties (such as validity, distortion and adequacy) that define tradeoffs designers must balance (Cardona-Rivera, 2020). Although those schemes are studied in a gaming context, the idea of balancing coverage against clarity is general: a coarse frame can handle many scenarios but may lack precision, whereas a fine-grained frame is more interpretable yet applies to fewer situations. Our design favours interpretability and human-aligned semantics while recognising that greater detail must be incorporated carefully.

2.2 Neuro-Symbolic AI for Scene Understanding

Neuro-symbolic AI (NSAI) seeks to integrate the learning capabilities of deep neural networks with the explicit reasoning of symbolic systems (Besold et al., 2017; d'Avila Garcez & Lamb, 2020). Neural models excel at perception and pattern recognition, whereas symbolic reasoning contributes transparency, compositional structure, and the ability to generalize across novel contexts. This integration has proved effective in diverse tasks including program synthesis (Chaudhuri, 2025) and visual question answering (Yi et al., 2018). In the context of autonomous driving, structured representations such as lane-topology graphs and scene graphs demonstrate how explicit structure improves interpretability and retrieval (Xu et al., 2017; Tang et al., 2020; Fu et al., 2024).

Within autonomous-driving research, NSAI approaches have been applied to enhance safety, policy interpretability, and explanatory reasoning. (Qi et al., 2024) integrate logic-based temporal constraints into neural controllers to ensure risk-aware decision making. (Sharifi et al., 2023) embed symbolic reinforcement-learning rules into deep networks to obtain verifiable driving policies. (Yuan et al., 2024) employ retrieval-augmented reasoning with large language models to produce structured, case-based explanations of driving behavior. Collectively, these studies exemplify complementary neuro-symbolic strategies that bridge perception and reasoning through formal logic, hybrid planning, and structured retrieval. Our work extends this line of research by deriving symbolic scene representations directly from sensory data and aligning them with cognitively motivated frame templates, thereby enabling interpretable and introspective scene understanding.

We process NuScenes (Caesar et al., 2020) logs into symbolic frame hierarchies, enriched with natural language descriptors and semantic roles. These representations enable context-aware behavior modeling, anomaly detection, and uncertainty reflection. Compared to prior works, our frames are dynamic, interpretable, and grounded in real-world motion data.

2.3 Language-Based Reasoning with Symbolic Abstractions

LLMs have recently demonstrated emergent capabilities in causal reasoning, narrative simulation, and physical commonsense inference (Huang et al., 2022; Kosinski, 2023; Bubeck et al., 2023). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) enhances these models by injecting structured context during inference. RAG methods have been adapted for driving policy explanation and decision support in autonomous driving(Yuan et al., 2024; Hussien et al., 2025).

We build on this paradigm by conditioning LLMs on our structured traffic frames. These inputs summarize motion, spatial intent, and agent interactions using symbolic slot-filler representations. Rather than responding to raw visual inputs, the language model reasons over abstracted cognitive representations, simulating mental models and producing reflective outputs.

This integration of symbolic abstraction with generative reasoning supports a research path toward self-aware and communicative AI agents in autonomous systems. In this work, we present a preliminary framework focused on the methodological design and theoretical grounding, with comprehensive empirical evaluation planned for future work.

3. Cognitive Perspective

Human scene understanding operates through multiple interconnected cognitive processes that transform sensory input into structured, actionable knowledge. Cognitive psychology has extensively documented these processes, from attention and pattern recognition to working memory integration and episodic storage (Barsalou, 2012). Frame Theory (Minsky, 1974) offers a computational framework that captures many of these mechanisms: stereotyped situations with slots (roles/attributes) and fillers (instantiations) that activate expectations, guide attention, and highlight anomalies when observations deviate from prior experience.

Our hierarchical frame design is explicitly aligned with established models of human cognition. Research in cognitive development shows that humans construct layered mental representations, from perceptual features to object concepts to situation models, that support flexible reasoning and transfer (Treisman & Gelade, 1980; Zwaan & Radvansky, 1998). Our frames are designed to mirror these core cognitive processes:

- **Expectation and default reasoning**: Human drivers rely heavily on schematic knowledge to predict typical behaviors (e.g., vehicles stopping at red lights, pedestrians yielding at crosswalks). Our slots encode these affordances and support completion of missing details when observations are incomplete or occluded (Barsalou, 2012). This mirrors schema-based processing in human memory, where prototypical knowledge fills gaps in perception.

- **Attention and anomaly detection**: Cognitive research shows that humans automatically detect violations of expected patterns, which trigger increased attention and processing resources (Minsky, 1974). Unfilled or conflicting slots in our frames serve this function, exposing knowledge gaps and behavioral violations that require deeper analysis or immediate action.

- **Episodic integration**: The progression `SampleFrame` → `ObjectFrame` → `SceneFrame` mirrors human temporal processing from momentary perceptual snapshots to object-centric situation models to episodic scene summaries. This hierarchical integration is fundamental to human event comprehension (Zwaan & Radvansky, 1998; Lake et al., 2017).

- **Metacognition and self-awareness**: Human cognition is distinguished by explicit awareness of knowledge states, including what we know, what we do not know, and what information is needed for confident action (Flavell, 1979; Schraw & Moshman, 1995). Our frames make epistemic states explicit through unfilled slots and confidence markers, enabling computational systems to reflect on perceptual sufficiency.

- **Mental simulation and counterfactual reasoning**: Research on human imagination shows that people construct internal models to simulate unseen events and explore alternative scenarios (Lake et al., 2017). By conditioning LLMs on symbolic frames rather than raw pixels, we enable similar narrative inference and counterfactual exploration based on structured situation models.

This cognitive alignment is intended to move autonomous systems beyond correlational pattern recognition toward more structured forms of understanding, explanation, and adaptive reasoning. The architecture embodies a principled division of cognitive labor reminiscent of dual-process theories in psychology: fast, automatic sub-symbolic modules perform efficient feature extraction and object recognition; structured frames organize these features into interpretable, queryable situation models; and higher-level language models operate on these symbolic representations to communicate, hypothesize, and reflect. This three-tier organization is consistent with theoretical accounts of hierarchical cognition and with broader calls for robust, interpretable AI architectures that integrate symbolic and sub-symbolic reasoning (Marcus, 2020; Eckstein & Collins, 2021).

4. Method

4.0.1 Dataset Schema and Terminology

Our system implements a frame-based abstraction pipeline for autonomous driving scenes using the NuScenes dataset (Caesar et al., 2020), one of the most comprehensive public benchmarks for autonomous vehicle perception and prediction. NuScenes consists of 1000 driving episodes collected in complex urban environments across Boston and Singapore, capturing diverse weather conditions, traffic densities, and road configurations. The dataset provides 3D bounding box annotations for 23 object categories, instance tracking across time, and high-definition semantic maps, making it ideal for studying multi-agent interactions and long-term behavior patterns in real-world traffic scenarios.

The NuScenes schema is composed of several key entities (Table 1):

Term	**Definition**
Scene	A scene is a continuous 20-second driving interval and the top-level container in the NuScenes dataset. It consists of a chronologically ordered sequence of `Samples`.
Sample	A sample represents a timestamped snapshot of the scene. It is recorded at 2 Hz and links to all associated sensor readings and annotations at that time.
SampleData	Each sample is composed of multiple sensor streams, including top-mounted lidar, cameras, and radar. These are stored as distinct `SampleData` entries.
Instance	An instance is a unique object (e.g., car, pedestrian, bicycle) tracked throughout one or more samples. Each instance has a persistent ID.
Annotation	An annotation describes the state of an instance at a particular sample, including its 3D bounding box, size, position, velocity, orientation, and category.
Ego Vehicle	The ego vehicle refers to the data-collection platform that all sensors are mounted on. All annotations are expressed in the ego-centric coordinate frame.
Location	A location refers to a high-definition map region (e.g., Boston Seaport or Singapore Onenorth). Map priors include road lanes, traffic rules, and spatial context.
Ego View (EV)	A dynamic coordinate system centered on the ego vehicle. All nearby object trajectories are computed relative to the ego's position and heading. Useful for reasoning about first-person interactions.
Bird's Eye View (BEV)	A ground-plane projection of the scene from above. Commonly used for interpreting spatial relationships between agents. Derived from the top-mounted lidar sensor.

Table 1. Core schema components and concepts from the NuScenes dataset used in our frame-based interpretation pipeline.

The top-mounted lidar (referred to as `LIDAR_TOP`) is the primary sensor used in our implementation for 3D localization and trajectory interpretation. Each lidar point cloud is transformed into a bird's-eye view (BEV) representation. The BEV is a projection of the environment onto the ground plane, allowing us to compute relative distances and movement patterns for all annotated agents.

4.0.2 Frame Abstraction and Design

To convert raw NuScenes data into structured semantic representations, we design a hierarchical abstraction pipeline grounded in Frame Theory (Minsky, 1974). This approach decomposes the continuous stream of sensor readings into cognitively meaningful units, organized across three conceptual levels: `sample_frame`, `object_frame`, and `scene_frame`. This design reflects how humans incrementally understand dynamic environments, from momentary observation to object behavior to scene-level inference.

Semantic Hierarchy Overview Each level in our system captures a distinct scope of interpretation:

- **`SampleFrame`**: Encodes the scene at an individual timestamp. It contains the calibrated ego pose, object annotations, and relative spatial relationships. This level mirrors the sensory snapshot an autonomous vehicle processes every 0.5 seconds.
- **`ObjectFrame`**: Compiles temporally ordered sample frames for each tracked instance (using the `instance_token`). It computes motion statistics, interaction types (e.g., overtaking, crossing), and symbolic behavior tags over time.
- **`SceneFrame`**: Summarizes all object frames within a driving episode (`scene`) and integrates high-level context (e.g., location, route summary, traffic configuration). This level supports holistic reasoning and scenario-based querying.

This hierarchy enables structured representations that are both temporally coherent and relationally grounded. Crucially, each frame is ego-aligned, meaning object states are expressed relative to the vehicle's perspective, which facilitates interaction analysis and commonsense interpretation across different driving contexts.

Motivation and Design Principles Our motivation stems from the need for interpretable, modular, and cognitively inspired representations in autonomous driving systems. Frame Theory offers a natural template for such representations. Each frame defines expected roles (slots), which are instantiated over time through sensor perception. When slots remain unfilled or deviate from expectations, they highlight uncertainty or anomalous behavior, which provides hooks for explanation and policy refinement.

The separation between frame levels enables scalable processing: `sample_frames` support low-latency updates, `object_frames` allow for mid-level summarization, and `scene_frames` capture episodic context. Each frame class exposes methods for rule-based updating, symbolic labeling, and natural language generation. This modularity is key for downstream integration with reasoning engines and large language models.

Frame Implementation in Code Our implementation follows a class-based structure in Python. Each frame is a class instance with typed slots, symbolic fields, and methods for assignment, update, and export. For example, the `ObjectFrame` includes routines to segment motion into directional primitives, detect behavior patterns (e.g., "crossed in front of ego"), and generate structured text (e.g., "A cyclist overtook the ego vehicle from the left between 4.0s and 7.0s").

Each frame instance is serialized into a nested dictionary and saved in JSON format for downstream use. This makes them accessible to both symbolic reasoning tools and neuro-symbolic models conditioned on language. Examples of frame instantiation are illustrated in Figure 1.

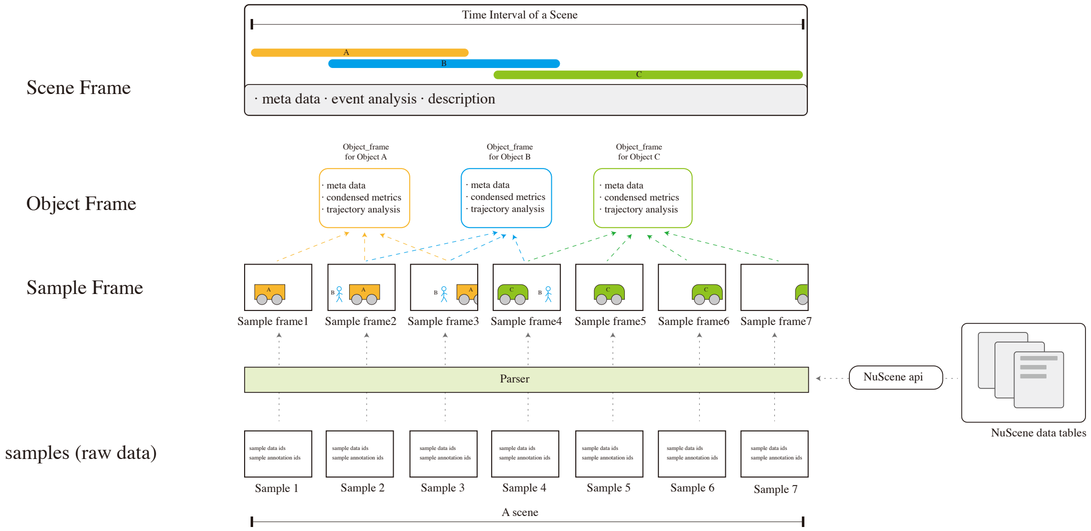


Figure 1. Three-level frame abstraction hierarchy: `SampleFrame` (timestamped, ego-aligned observations), `ObjectFrame` (temporal aggregation and behavior labels), and `SceneFrame` (episode-level composition and context).

4.0.3 Annotation and Instance Tracking

Our annotation and tracking pipeline consists of three stages: (1) metadata extraction from raw NuScenes tables, (2) ego-centric calibration and instance trajectory accumulation, and (3) rule-based interpretation and symbolic population.

(1) Metadata Extraction from Raw Sensor Tables We traverse each `scene`, follow its linked chain of `sample` tokens at 2 Hz, retrieve the associated `sample_data` entries, and record the `ego_pose` per timestamp. For each sample, we pull all 3D bounding-box `annotation` entries, including position, size, orientation, velocity (when available), and semantic category. This yields synchronized time series for ego and surrounding instances.

(2) Ego-Centric Calibration and Instance Accumulation We transform all object coordinates into the ego-centric frame using the inverse of the `ego_pose` homogeneous transform. This normalization expresses trajectories relative to the ego vehicle's position and yaw, which supports interaction analysis. We then group annotations by `instance_token` to build ordered trajectories of time-stamped tuples (x, y, θ), forming the input to `ObjectFrame` construction. A synchronized ego pose log is likewise maintained to derive relative dynamics such as overtaking, following, and crossing.

(3) Rule-Based Interpretation and Symbolic Population From calibrated trajectories, we compute first- and second-order motion statistics (stepwise displacement, speed profile, lateral shifts, heading changes), then apply a set of rules to detect behavioral primitives. For instance, a sign change in lateral position near the ego origin indicates a potential "crossed in front" event, while sustained parallel motion with decreasing lateral offset suggests an overtake.

4.0.4 Ego Motion Classification Rules

We segment the ego trajectory into symbolic motion primitives using change points in movement and orientation. Two core rulesets handle forward/backward detection and lateral motion classification.

When the angle between the ego orientation and movement vectors is less than $\pi/2$, motion is labeled forward; otherwise, it is labeled as reversing. This classification disambiguates turning semantics in subsequent steps. Let $\Delta\theta$ be the change in movement angle between successive timesteps. We use thresholds of 2°, 5°, and 10° to label segments as straight, drifting, turning, or sharp turning, with direction determined by the sign of $\Delta\theta$ and adjusted for forward/back status.

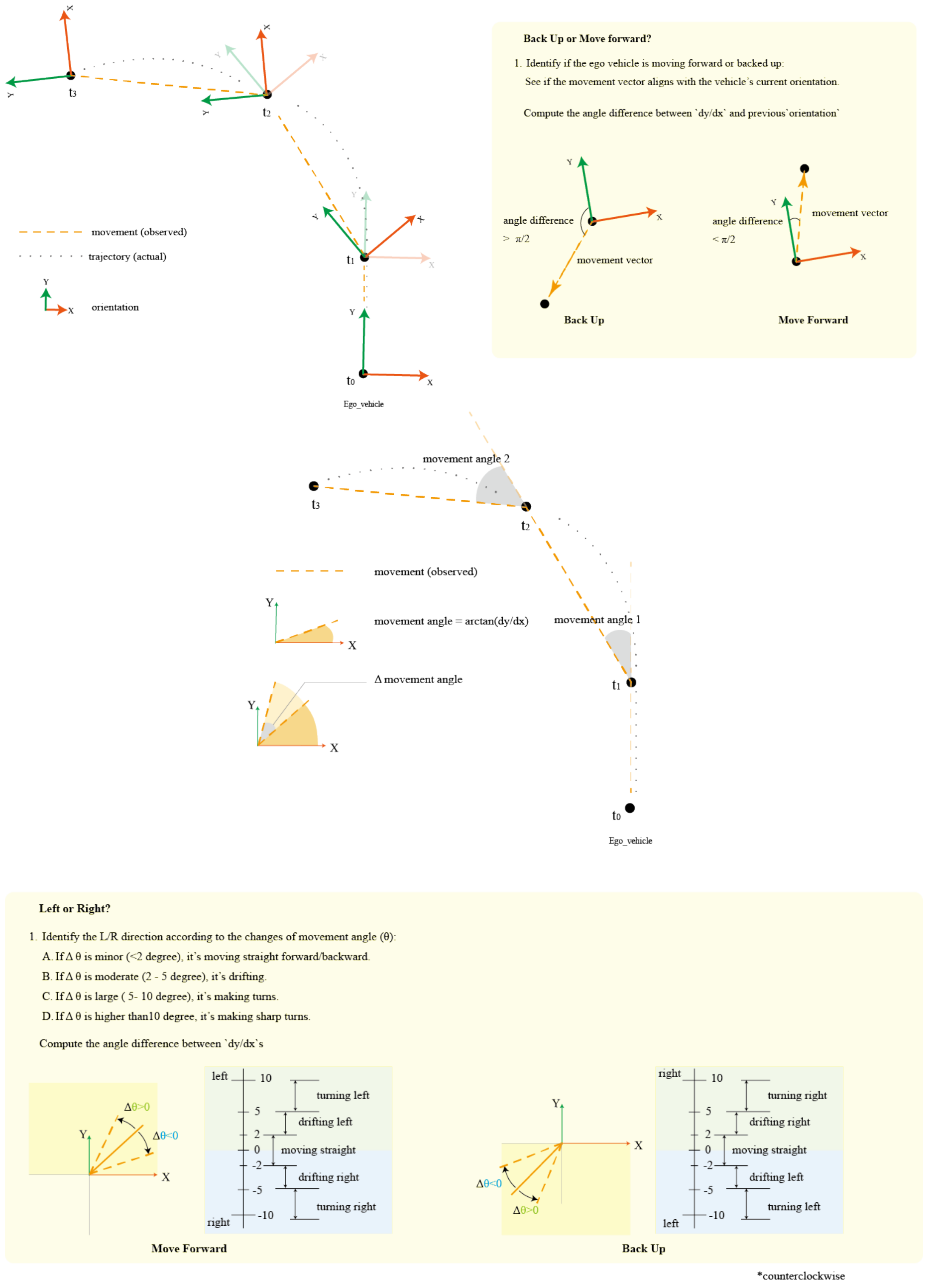
4.0.5 Directional Segmentation for Instances

For each `ObjectFrame`, we segment movement by dominant direction using the ratio $r = |\Delta x/\Delta y|$ and a minimum step length to filter minor motion. We track side transitions (left/right, ahead/behind), orientation classes, and group adjacent segments.

Algorithm 1: Instance Trajectory Interpretation via Directional Segmentation

```
Input: Ego-aligned object trajectory: $(x_i, y_i)$ with timestamps
Output: Symbolic movement summary for the instance
Initialize empty segment list and relative position trackers;
foreach consecutive position pair $(x_i, y_i), (x_{i+1}, y_{i+1})$ do
    Compute $\Delta x$, $\Delta y$, and distance;
    if distance $< 0.5$ m then
        label: minor movement
    else
        set $r = |\Delta x/\Delta y|$ and label as dominantly left/right if $r > 1.5$, dominantly
          forward/back if $r < 0.66$, otherwise diagonal
    Update lateral side (left/right) and longitudinal side (ahead/behind);
    Update orientation class from yaw; group segments and timestamps accordingly;
Summarize episodes, position changes, distances, and temporal span.
```

Robustness and Complexity Real-world trajectory data from lidar and tracking systems contains noise, quantization artifacts, and occasional tracking failures that can lead to spurious motion labels. To ensure robust segmentation, we implement several filtering strategies: (i) a 3-point temporal median filter on (x, y) coordinates before computing differences, which eliminates single-frame outliers while preserving genuine motion changes; (ii) hysteresis thresholds that require sustained change before switching labels, preventing rapid oscillation between categories; and (iii) a post-



Back Up or Move forward?

1. Identify if the ego vehicle is moving forward or backed up:
   See if the movement vector aligns with the vehicle's current orientation.

Compute the angle difference between `dy/dx` and previous`orientation`

Left or Right?

1. Identify the L/R direction according to the changes of movement angle (θ):
   A. If Δ θ is minor (<2 degree), it's moving straight forward/backward.
   B. If Δ θ is moderate (2 - 5 degree), it's drifting.
   C. If Δ θ is large (5- 10 degree), it's making turns.
   D. If Δ θ is higher than10 degree, it's making sharp turns.

Compute the angle difference between `dy/dx`s

Figure 2. Top: determining forward or backward motion by comparing the orientation vector (green) with the movement vector (orange). Bottom: classifying lateral motion using change in movement angle $\Delta\theta$ into straight, drifting, turning, or sharp turning, with left/right decided by the sign.

processing step that merges adjacent segments with identical labels separated by gaps shorter than

0.5 seconds, which accounts for brief tracking interruptions or minor motion below the detection threshold.

The minimum segment duration is set to 0.5–1.0 seconds to focus on behaviorally meaningful motion patterns rather than micro-adjustments.

Algorithm 2: Ego Vehicle Trajectory Interpretation

```
Input: Ego position and yaw sequences over time
Output: Symbolic motion segments and aggregated statistics
Sort by time; initialize totals and previous pose;
foreach consecutive pose pair do
    Compute movement vector and distance; movement and yaw angle changes; normalize
      to $[-\pi, \pi]$;
    Determine forward/back via angle between movement and orientation;
    Label as straight/drift/turn/sharp turn with left/right;
    Append to segment list and accumulate distance;
Group identical labels into contiguous segments with timestamps; compute duration,
  average speed, and displacement.
```

For ego segmentation we additionally unroll yaw angles to avoid wrapping artifacts, and we snap tiny $|\Delta\theta|$ to zero under a tolerance to improve stability. A post-pass merges micro-segments under a duration threshold and re-computes summary statistics.

4.0.6 Scene-Level Composition and Language Summaries

We compose `SceneFrame` by aggregating all `ObjectFrames`, ego trajectory segments, and global context (location, weather if available), producing: (i) episodic statistics (duration, object counts, interaction counts), (ii) salient events (e.g., "pedestrian crossed ahead during ego right turn"), and (iii) a natural language summary. Summaries are generated from templates populated by symbolic slots and can be refined via LLM paraphrasing for fluency.

4.0.7 Hyperparameters and Implementation Details

By default, we use a sampling rate of 2 Hz, a minor movement threshold of 0.5 m, direction-dominance ratio thresholds of 0.66 and 1.5, heading-change thresholds of 2°, 5°, and 10°, and a 100–200 m BEV region of interest centered on the ego for visualization. All frames are serialized as nested JSON files for downstream querying and retrieval. The current thresholds were chosen empirically to balance sensor precision and interpretability. A 0.5 m cutoff filters typical lidar noise (approximately 0.3 m) while retaining meaningful displacement, and angular limits of 2°, 5°, and 10° were found to reasonably separate noise, gentle drift, and deliberate turns. These settings produced stable segmentation across representative scenes, though they remain adjustable and may be refined in future work.

5. Visualization

Effective visualization is crucial for validating frame abstractions, debugging algorithmic decisions, and communicating results to human annotators. We provide three complementary visualization modes, each designed to highlight different aspects of the hierarchical frame structure and support different analytical tasks.

- **Instance-level trajectory analysis**: This mode renders ego and selected object trajectories with time-gradient coloring (early positions in light colors, recent positions in saturated hues), directional heading ticks every 0.5 seconds, and class-specific geometric markers (rectangles for vehicles, circles for pedestrians, diamonds for cyclists). Object saliency for display selection is computed using a weighted combination of proximity to ego (inverse distance), temporal duration (longer presence increases salience), and interaction significance (crossing, overtaking, and yielding events receive higher weights). This view enables detailed analysis of individual object behaviors and their relationships to ego motion.
- **Scene map overlays with semantic context**: This mode plots the top-k most salient trajectories over high-definition NuScenes semantic maps, providing crucial spatial context including lane boundaries, intersection geometry, and traffic control devices. A red dashed region-of-interest (ROI) box spanning 100–200 meters is centered on the ego vehicle to focus attention on immediate interactions. Per-segment symbolic labels for ego motion primitives ("straight," "drifting left," "right turn") are placed at temporal anchors to show the correspondence between geometric motion and symbolic interpretation. This map facilitates the evaluation process of the semantic features extracted with our framework.
- **Temporal visualization and replay**: Dynamic visualizations replay scenes at the original 2 Hz frequency with alpha-blended trails that fade older positions to convey motion history and temporal flow. Symbolic behavior labels appear at appropriate anchor times with smooth transitions and remain visible for sufficient duration to be readable. Key interaction events (crossing initiation, overtaking completion) are highlighted with distinctive markers and brief text annotations. Animations are exported as high-quality GIFs to support human evaluation.

We follow three visualization principles. First, we use high-contrast color schemes and a consistent ego-centric coordinate system to improve readability and comparability across scenes. Second, we place text with a non-maximum suppression strategy that reduces label overlap and preserves local context. Third, we adapt rendering density by subsampling in long or cluttered episodes while preserving anchor events and state transitions.

We will extend these visualizations to (i) support side-by-side comparison between rule-based labels and LLM predictions for each `ObjectFrame` and `SceneFrame`, (ii) highlight disagreements with color-coded overlays and timestamps to localize errors, and (iii) export per-scene summary sheets that align textual narratives with spatial plots. These tools will facilitate expert review of the evaluation protocol and help diagnose failure modes such as missed interactions or inconsistent temporal ordering.

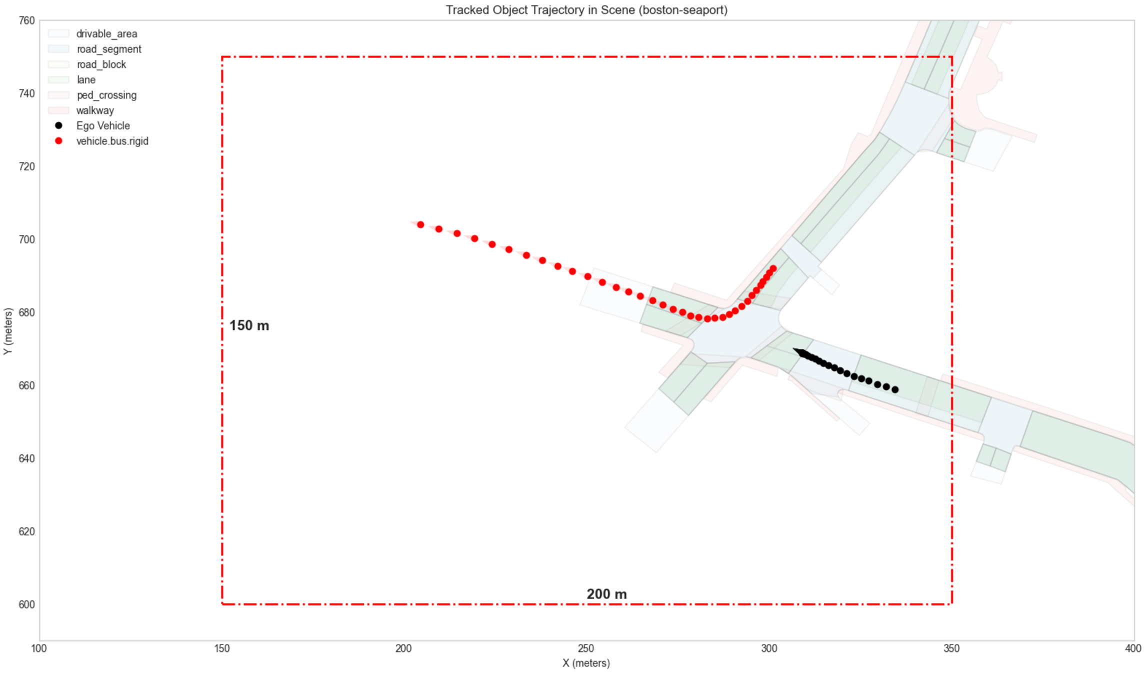


Figure 3. Instance-level visualization showing ego and bus trajectory interaction.

6. Language-Based Reasoning

A key hypothesis of our work is that large language models can perform structured reasoning over symbolic scene representations without direct access to raw sensory data. This capability would enable hybrid systems that combine high-performance perception with human-like semantic reasoning and communication. We test this hypothesis by interfacing frame-derived JSON summaries with carefully designed LLM prompts.

Recent studies indicate that modern LLM/VLM systems can coordinate multi-step spatial reasoning and integrate symbolic context when properly conditioned (Wen et al., 2023; Chen et al., 2023; Liu et al., 2023a; Sima et al., 2024). Our contribution is to ground these abilities in real driving scenarios through cognitively motivated frames. We hypothesize that conditioning on structured frame representations, rather than pixels or free-form text, yields more accurate and introspectively aware reasoning about traffic scenes.

Evaluation Protocol We use labels produced by our rule-based hierarchical frames as ground truth to probe two capabilities of language models.

1. **Label induction from low-level inputs**. Given low-level spatial inputs (ego-relative coordinates and pairwise relations), can the model infer the same frame-consistent labels as our rules at the `ObjectFrame` and `SceneFrame` levels (e.g., motion segments, interaction labels).

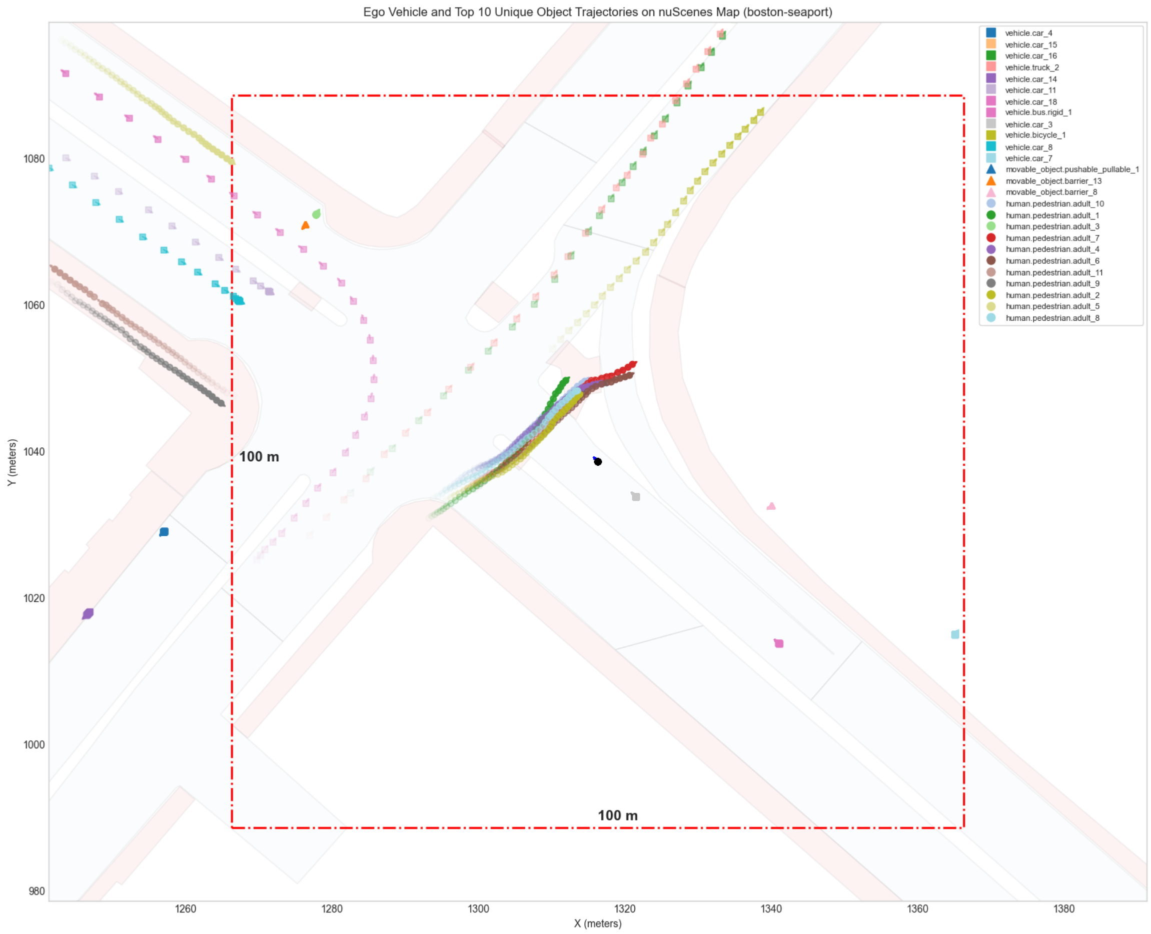


Figure 4. Map visualization of ego and salient object trajectories.

2. **Scene summary from labels**. Given only the symbolic labels, can the model generate a coherent scene summary that preserves spatial relations, temporal order, and salient interactions, consistent with `SceneFrame` narratives.

We use a single, structured prompt aligned with the frame design to elicit zero-shot, single-turn responses for both tasks. The prompt contains: (i) an ego-vehicle summary with temporal motion segments and aggregate statistics; (ii) the top-k salient objects with their behavioral labels, proximity measures, and interaction classifications; (iii) concise, timestamped interaction snippets for key events (crossings, overtakes, yields); and (iv) a section listing unfilled or uncertain slots to make epistemic limitations explicit. This design reflects our cognitive framing: perceptual samples are abstracted into symbolic labels that support human-like understanding and are directly communicable to language models.

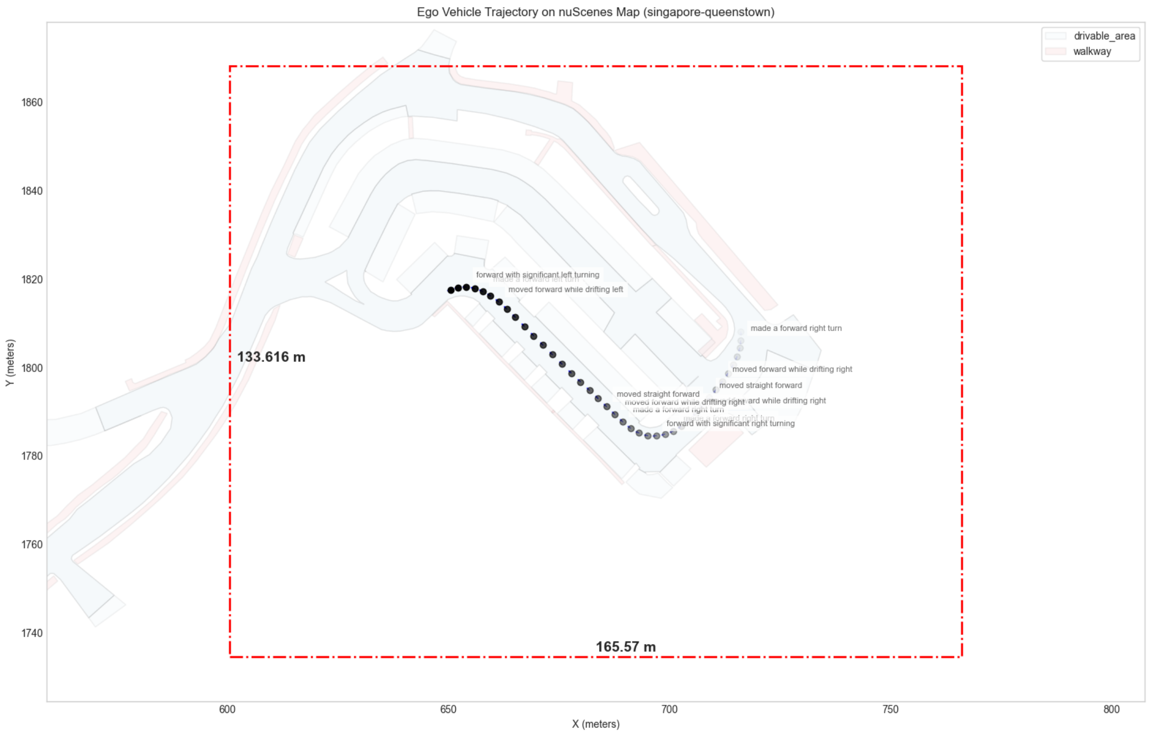

Ego Vehicle Trajectory on nuScenes Map (singapore-queenstown)
drivable_area
walkway
1860
1840
1820
1800
1780
1760
1740
Y (meters)
600
650
700
750
800
X (meters)
133.616 m
165.57 m
forward with significant left turning
made a forward left turn
moved forward while drifting left
made a forward right turn
moved forward while drifting right
moved straight forward
moved forward while drifting right
moved straight forward
moved forward while drifting right
made a forward right turn
made a forward right turn
forward with significant right turning

Figure 5. Ego vehicle movement segmentation for trajectory dynamics.

7. Limitations and Future Work

The current rule-based parsing operates in a simplified, abstract setting. It can over-segment curved motion, lacks global turn awareness, and may miss corner cases that violate template assumptions. Our narrative generation also relies on fixed templates, which limits verbal diversity and complicates evaluation because both references and LLM outputs are textual. Next, we will learn turn-aware primitives and induce rules from data, enrich frames with a prior-knowledge base (traffic norms, map/topology constraints, commonsense relations), and integrate planning/safety feedback. We also plan retrieval and case-based reasoning over stored `SceneFrames`, and human studies to assess explanation quality and trust. Finally, we will test portability across datasets and domains and explore automatic metrics that align summaries with symbolic constraints to enable fair comparison between LLM outputs and reference narratives. We will release our rule-derived frame labels (with JSON schemas and utilities) as a benchmark for reproducible LLM-based scene reasoning.

Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0149.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Barsalou, L. W. (2012). Frames, concepts, and conceptual fields. *Frames, fields, and contrasts*, (pp. 21–74).

Besold, T. R., et al. (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*.

Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Caesar, H., et al. (2020). nuscenes: A multimodal dataset for autonomous driving. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11621–11631).

Cao, J., Pang, J., Weng, X., Khirodkar, R., & Kitani, K. (2023). Observation-centric sort: Rethinking sort for robust multi-object tracking. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9686–9696).

Cardona-Rivera, R. (2020). Foundations of a computational science of game design: Abstractions and tradeoffs. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 167–174).

Chai, Y., Sapp, B., Bansal, M., & Anguelov, D. (2019). Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*.

Chaudhuri, S. (2025). Neurosymbolic program synthesis. In *Handbook on neurosymbolic ai and knowledge graphs*, 532–549. IOS Press.

Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., & Liu, Z. (2023). Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, *36*, 70115–70140.

Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., & Li, H. (2024). End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1290–1299).

Chittaro, L., Guida, G., Tasso, C., & Toppano, E. (1993). Functional and teleological knowledge in the multimodeling approach for reasoning about physical systems: a case study in diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*, 1718–1751.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Du, J., Su, S., Fan, R., & Chen, Q. (2023). Bird's eye view perception for autonomous driving. *Autonomous Driving Perception: Fundamentals and Applications*, (pp. 323–356).

Eckstein, M. K., & Collins, A. G. (2021). How the mind creates structure: Hierarchical learning of action sequences. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*, 906–11.

Fu, Y., Liao, W., Liu, X., Xu, H., Ma, Y., Zhang, Y., & Dai, F. (2024). Topologic: An interpretable pipeline for lane topology reasoning on driving scenes. *Advances in Neural Information Processing Systems*, *37*, 61658–61676.

d'Avila Garcez, A., & Lamb, L. C. (2020). Neurosymbolic ai: the 3rd wave. *arXiv e-prints*, (pp. arXiv–2012).

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE conference on computer vision and pattern recognition* (pp. 3354–3361). IEEE.

Huang, W., et al. (2022). Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

Hussien, M. M., Melo, A. N., Ballardini, A. L., Maldonado, C. S., Izquierdo, R., & Sotelo, M. A. (2025). Rag-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models. *Expert Systems with Applications*, *265*, 125914.

Karnchanachari, N., et al. (2024). Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 629–636). IEEE.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, *4*, 169.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 9459–9474).

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L. M., & Shum, H.-Y. (2023). Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3041–3050).

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., & Dai, J. (2024). Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liang, C., Berant, J., Le, Q., Forbus, K. D., & Lao, N. (2016). Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. *arXiv preprint arXiv:1611.00020*.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*, 31–57.

Liu, F., Emerson, G., & Collier, N. (2023a). Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, *11*, 635–651.

Liu, H., Teng, Y., Lu, T., Wang, H., & Wang, L. (2023b). Sparsebev: High-performance sparse 3d object detection from multi-camera videos. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 18580–18590).

Marcus, G. (2020). The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

McCarthy, J., & Hayes, P. J. (1981). Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, 431–450. Elsevier.

Minsky, M. (1974). A framework for representing knowledge.

Qi, S., Zhang, Z., Sun, Z., & Haesaert, S. (2024). Risk-aware autonomous driving with linear temporal logic specifications. *arXiv preprint arXiv:2409.09769*.

Salzmann, T., Ivanovic, B., Chakravarty, P., & Pavone, M. (2020). Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control. *arXiv preprint arXiv:2001.03093*, *2*, 1.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, *7*, 351–371.

Seff, A., Cera, B., Chen, D., Ng, M., Zhou, A., Nayakanti, N., Refaat, K. S., Al-Rfou, R., & Sapp, B. (2023). Motionlm: Multi-agent motion forecasting as language modeling. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8579–8590).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Sharifi, I., Yildirim, M., & Fallah, S. (2023). Towards safe autonomous driving policies using a neuro-symbolic deep reinforcement learning approach. *arXiv preprint arXiv:2307.01316*.

Shi, J., Chen, J., Wang, Y., Sun, L., Liu, C., Xiong, W., & Wo, T. (2025). Motion forecasting for autonomous vehicles: a survey. *arXiv preprint arXiv:2502.08664*.

Shi, S., Jiang, L., Dai, D., & Schiele, B. (2022). Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, *35*, 6531–6543.

Sima, C., et al. (2024). Drivelm: Driving with graph visual question answering. *European conference on computer vision* (pp. 256–274). Springer.

Tang, K., Zhang, H., & Wu, B. (2020). Unbiased scene graph generation from biased training. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 3716–3725).

Tian, Z., Cui, J., Jiang, L., Qi, X., Lai, X., Chen, Y., Liu, S., & Jia, J. (2023). Learning context-aware classifier for semantic segmentation. *Proceedings of the AAAI conference on artificial intelligence* (pp. 2438–2446).

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*, 97–136.

Wen, L., et al. (2023). On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*.

Wilensky, R. (1987). Some problems and proposals for knowledge representation. memorandum ucb/csd 87/351. *University of California, Berkeley Electronic Research Laboratory*.

Wilson, B., et al. (2023). Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*.

Xu, D., Zhu, Y., Choy, C. B., & Fei-Fei, L. (2017). Scene graph generation by iterative message passing. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5410–5419).

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 1039–1050).

Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., & Gadd, M. (2024). Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2022). Bytetrack: Multi-object tracking by associating every detection box. *European conference on computer vision* (pp. 1–21). Springer.

Zhou, Z., Ye, L., Wang, J., Wu, K., & Lu, K. (2022). Hivt: Hierarchical vector transformer for multi-agent motion prediction. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8823–8833).

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, *123*, 162.