Planning Beyond Perception: Benchmarking LLM- and VLM-Based Reasoning for Autonomous Driving

Anonymous submission

Abstract

Recent advances in large multimodal foundation models (LLMs, VLMs, and VLAMs) have demonstrated promising capabilities in perception and reasoning across visual and linguistic modalities. Yet, their effectiveness in *long-horizon, safety-critical planning*—a core requirement for autonomous driving—remains insufficiently understood. This work presents **Planning Beyond Perception (PBP)**, a benchmark for systematically evaluating the planning and decision-making abilities of multimodal foundation models in realistic driving contexts. PBP encompasses tasks requiring **situational reasoning** under multimodal inputs, **plan decomposition and adaptation** across dynamic traffic scenarios, and **safety-aware control** constrained by real-world driving rules. Using standardized environments derived from CARLA and nuScenes, we assess multiple architectures, including LLM-, VLM-, and VLAM-based agents, on their ability to generate interpretable, robust, and executable driving plans. Our findings reveal that while these models excel in short-horizon perception and description, they exhibit significant limitations in causal reasoning, temporal abstraction, and reliable action synthesis. PBP provides an open, reproducible framework to benchmark and advance the development of *foundation models for trustworthy autonomous planning*.

Introduction

Autonomous driving (AD) integrates perception, prediction, planning, and control into a unified decision-making pipeline. While recent advances in perception and motion prediction have greatly improved environmental understanding, robust and interpretable *long-horizon planning* under uncertainty remains an open challenge. Conventional modular approaches—where perception, prediction, and control are optimized separately—often lack generalization and fail to handle unseen, multi-agent scenarios (Shalev-Shwartz, Shammah, and Shashua 2017; Chen et al. 2021).

In parallel, large multimodal foundation models (FMs), including large language models (LLMs) and vision-language(-action) models (VLMs, VLAMs), have shown remarkable generalization and reasoning ability across diverse domains (OpenAI 2023; Li et al. 2023; Driess et al. 2023; Brohan et al. 2023). Recent work such as DriveLM (Chen et al. 2023) and Talk2Drive (Mehta et al. 2024) demonstrates that foundation models can connect perception with linguistic reasoning and interactive decision making. Yet, despite these promising results, the capacity of FMs for *safety-critical, causal, and temporally consistent planning* in realistic driving contexts remains poorly characterized (Wang et al. 2024; Zeng, Zhang, and Wang 2024). To address this gap, we introduce **Planning Beyond Perception (PBP)**, a benchmark designed to evaluate the reasoning and planning capabilities of LLM-, VLM-, and VLAM-based agents for autonomous driving. PBP focuses on three core dimensions: (1) **situational reasoning** under multimodal sensory and map inputs, (2) **plan decomposition and adaptation** to dynamic traffic conditions, and (3) **safety-aware decision making** under real-world constraints. Using standardized environments derived from CARLA (Dosovitskiy et al. 2017) and nuScenes (Caesar et al. 2020), PBP measures plan validity, robustness under perturbation, and alignment with human rationales. By releasing this benchmark, task suite, and evaluation toolkit, our goal is to establish a reproducible and extensible foundation for *benchmarking planning-oriented reasoning* in multimodal foundation models—bridging perception, abstraction, and control in autonomous systems.

Problem Definition

We consider the problem of **autonomous driving planning with multimodal foundation models**. Let $\mathcal{E}$ denote the driving environment composed of a set of dynamic agents (e.g., vehicles, pedestrians) and static map elements (e.g., lanes, traffic lights, intersections). At each discrete timestep t, the ego vehicle observes multimodal sensory inputs:

$$\mathbf{o}_t = \{\mathbf{I}_t^{\text{cam}}, \mathbf{L}_t^{\text{lidar}}, \mathbf{M}^{\text{map}}, \mathbf{S}_t^{\text{state}}\},$$

where $\mathbf{I}_t^{\text{cam}}$ denotes camera images, $\mathbf{L}_t^{\text{lidar}}$ represents LiDAR point clouds, $\mathbf{M}^{\text{map}}$ provides high-definition maps or semantic lanes, and $\mathbf{S}_t^{\text{state}}$ encodes the ego state (position, velocity, heading).

A multimodal foundation model f_θ (e.g., LLM, VLM, or VLAM) conditions on these inputs and produces an interpretable plan representation:

$$\pi_t = f_\theta(\mathbf{o}_{1:t}) \in \mathcal{A}^H,$$

where π_t denotes a sequence of high-level or low-level actions (e.g., *turn-left, accelerate, yield*), and H is the planning horizon. The objective is to generate a plan that satisfies the following properties:

- **Executability** — the plan must be physically feasible within $\mathcal{E}$ and obey vehicle dynamics.
- **Safety** — the plan must remain collision-free and comply with traffic rules and environmental constraints.
- **Rationality** — the plan should align with human driving preferences and demonstrate causal reasoning.
- **Adaptivity** — the plan must remain robust under dynamic agent behavior and sensor perturbations.

Formally, the benchmark evaluates f_θ by minimizing a composite error functional:

$$\mathcal{L}_{\text{PBP}} = \lambda_1 \mathcal{L}_{\text{validity}} + \lambda_2 \mathcal{L}_{\text{safety}} + \lambda_3 \mathcal{L}_{\text{rationality}} + \lambda_4 \mathcal{L}_{\text{efficiency}},$$

where λ_i are weighting coefficients balancing different aspects of plan quality. Each term corresponds to metrics computed from simulation or human-annotated trajectories.

The overall goal of **Planning Beyond Perception (PBP)** is therefore to assess how well multimodal foundation models can transform raw sensory streams into interpretable, safe, and temporally consistent driving plans—bridging perception, reasoning, and control within a unified evaluation framework.

Proposed Benchmark

The **Planning Beyond Perception (PBP)** benchmark provides a standardized framework for evaluating multimodal foundation models (FMs) in *long-horizon, safety-critical planning*. Unlike prior driving benchmarks that focus primarily on perception or motion prediction, PBP explicitly measures the ability of models to reason, decompose, and execute structured driving plans from raw sensory input. The benchmark is designed around three complementary components: scenario suite, model interface, and evaluation metrics.

Scenario Suite

PBP includes a curated set of urban driving tasks derived from CARLA (Dosovitskiy et al. 2017) and nuScenes (Caesar et al. 2020). Each scenario $\mathcal{S}_i$ is represented as a tuple:

$$\mathcal{S}_i = (\mathbf{O}_{1:T}, \mathbf{A}^*_{1:T}, \mathbf{M}, \mathcal{C}),$$

where $\mathbf{O}_{1:T}$ denotes multimodal observations (camera, LiDAR, and map), $\mathbf{A}^*_{1:T}$ are ground-truth expert trajectories, $\mathbf{M}$ encodes the road topology and semantic map, and $\mathcal{C}$ represents environmental constraints such as traffic rules and safety zones. The benchmark covers diverse interaction-heavy scenes including intersections, lane changes, merges, and unprotected turns.

Model Interface

Each foundation model f_θ (LLM, VLM, or VLAM) receives sequential multimodal inputs and produces a structured plan $\pi_t = f_\theta(\mathbf{o}_{1:t})$. To ensure comparability, all models interact with the environment through a unified *perception-to-action API*:

$$\mathbf{a}_t = g(\pi_t, \mathbf{o}_t), \quad \text{where } \mathbf{a}_t \in \mathcal{A}$$

and $g(\cdot)$ maps high-level textual or symbolic plans into low-level control signals (steering, acceleration, braking). Optional modules allow planner-in-the-loop integration with A*, MCTS, or learned motion controllers for hybrid reasoning experiments.

Evaluation Metrics

To quantify overall planning quality, PBP defines a composite evaluation loss:

$$\mathcal{L}_{\text{PBP}} = \lambda_1 \mathcal{L}_{\text{validity}} + \lambda_2 \mathcal{L}_{\text{safety}} + \lambda_3 \mathcal{L}_{\text{rationality}} + \lambda_4 \mathcal{L}_{\text{robustness}},$$

where each component reflects a distinct aspect of planning performance:

- $\mathcal{L}_{\text{validity}}$: trajectory deviation from expert demonstrations or reachable goal states;
- $\mathcal{L}_{\text{safety}}$: penalty for collisions, rule violations, or unsafe distances;
- $\mathcal{L}_{\text{rationality}}$: divergence from human-preferred maneuvers and causal consistency checks;
- $\mathcal{L}_{\text{robustness}}$: stability under sensor noise or dynamic agent perturbations.

Each metric is normalized to $[0, 1]$ for fair cross-model comparison. The final benchmark score is defined as:

$$\text{Score}_{\text{PBP}} = 1 - \mathcal{L}_{\text{PBP}},$$

representing overall plan quality, where higher values indicate better performance.

Benchmark Goals

Through this formulation, PBP aims to (i) provide a reproducible testbed for evaluating multimodal reasoning-to-action capabilities, (ii) reveal failure modes in causal or temporal reasoning of foundation models, and (iii) support the development of interpretable, safety-aware agents that generalize across driving domains. The benchmark implementation, evaluation toolkit, and baseline model configurations are released to facilitate transparent and comparable research in *foundation-model-based autonomous planning*.

Expected Results

We expect multimodal foundation models to exhibit strong short-term situational reasoning but limited long-horizon consistency in autonomous planning. VLM-based agents are likely to outperform purely textual LLMs in spatial understanding, while VLAM architectures (e.g., PaLM-E (Driess et al. 2023), RT-2 (Brohan et al. 2023)) may achieve higher plan executability through grounded action representation. Nonetheless, challenges in causal dependency tracking, temporal abstraction, and adherence to safety constraints are anticipated to remain open problems.

Through this benchmark, we aim to reveal the trade-off between multimodal grounding and planning reliability, offer diagnostic insights into reasoning bottlenecks, and establish baseline performance for reproducible, trustworthy autonomous decision-making. In summary, **Planning Beyond Perception (PBP)** provides a unified framework for evaluating and advancing *foundation-model-based planning*—bridging perception, reasoning, and control toward safer and more interpretable autonomous systems.

References

Brohan, A.; Brown, N.; Chebotar, Y.; and et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *arXiv preprint arXiv:2307.15818*.

Caesar, H.; Bankiti, V.; Lang, A. H.; and et al. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11621–11631.

Chen, D.; Zhou, B.; Koltun, V.; and Krähenbühl, P. 2021. Learning by Cheating: Using Simulation to Train Real-World Autonomous Driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1239–1248.

Chen, X.; Xu, H.; Liu, H.; et al. 2023. DriveLM: Driving with Large Language Models. *arXiv preprint arXiv:2312.05832*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator. In *Conference on Robot Learning (CoRL)*, 1–16.

Driess, D.; Xia, F.; Sajjadi, M. S.; and et al. 2023. PaLM-E: An Embodied Multimodal Language Model. In *arXiv preprint arXiv:2303.03378*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Mehta, Y.; Bhatnagar, S.; Agarwal, R.; et al. 2024. Talk2Drive: Language-Guided Driving with Multimodal Feedback. *arXiv preprint arXiv:2403.01234*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Shalev-Shwartz, S.; Shammah, S.; and Shashua, A. 2017. On a Formal Model of Safe and Scalable Self-Driving Cars. *arXiv preprint arXiv:1708.06374*.

Wang, X.; Wang, L.; Wu, Z.; and Zhao, H. 2024. LLMs Meet Autonomous Driving: A Survey and Perspective. *arXiv preprint arXiv:2403.05750*.

Zeng, X.; Zhang, Y.; and Wang, J. 2024. Language-Driven Simulation: LLM-Based Scenario Generation for Safety Testing of Autonomous Vehicles. *IEEE Transactions on Intelligent Vehicles*, 9(4): 712–726.