Master's Thesis Proposal

Vision-Only Planning Transformer for Unified Trajectory Planning and Control in Autonomous Driving

Matthew Evans
The University of Texas at Dallas
matthew.evans@utdallas.edu

May 14, 2025

Abstract

We propose a Vision-Only Planning Transformer for end-to-end autonomous driving that will unify long-horizon trajectory planning and low-level vehicle control within a single model. Our architecture will process sequences of monocular camera frames and vehicle state, interleaving learnable Planning Tokens—coarse intent representations predicted at fixed intervals—with standard frame tokens. These Planning Tokens will guide subsequent control outputs, mitigating compounding error common in autoregressive policies. We will train the model via imitation learning on CARLA-generated urban driving data to predict both high-level waypoints and instantaneous steering, throttle, and brake commands. In closed-loop evaluation, we expect the Planning-Token–enhanced Transformer to outperform a baseline vision-only model in route completion, trajectory accuracy, and collision avoidance. We also expect attention visualizations to reveal that Planning Tokens focus on distant road features while control tokens attend to immediate obstacles. This work will demonstrate the efficacy of implicit hierarchical planning in a unified vision-only framework.

Introduction

This proposal presents a Vision-Only Planning Transformer that will unify long-horizon trajectory planning and low-level control within a single end-to-end model. The proposed architecture will process a history of monocular frames and vehicle state, will interleave learnable Planning Tokens—coarse intent representations at fixed intervals—with frame and state tokens, and will output both future waypoints and instantaneous control commands. This dual time-scale prediction is designed to mitigate the compounding error common in purely autoregressive policies [1].

Key contributions include:

- **Unified Architecture:** A single Transformer that jointly performs planning and control, eliminating separate modules.
- **Planning Tokens:** High-level intent tokens that will guide long-term behavior without sacrificing responsiveness to immediate scenarios.

- **Reproducibility:** A research prototype using a single front-facing camera and limited compute resources (4× A30 or H100), accompanied by code, trained models, and simulation results.

The model will be evaluated in CARLA [2], where we expect it will achieve higher route completion rates, lower trajectory deviation, and reduced collision frequency compared to a baseline without Planning Tokens. We also anticipate that attention visualizations will demonstrate that Planning Tokens attend to distant landmarks (e.g., upcoming intersections), while control tokens focus on immediate hazards.

Background and Related Work

Classical autonomous driving stacks decompose the pipeline into perception, prediction, planning, and control modules. While modularity aids interpretability and safety validation, inter-module errors accumulate and require extensive hand-tuning [3, 4]. Early end-to-end approaches—e.g., ALVINN [5] and NVIDIA's PilotNet [6]—map raw images directly to steering commands, reducing engineering effort but often failing to generalize due to short-horizon reasoning.

We will build upon recent Transformer-based models for spatiotemporal driving tasks. For example, TransFuser [7] fuses multi-sensor features for mid-level affordance and trajectory prediction, and ChauffeurNet [8] learns end-to-end vision-based driving policies via imitation. However, these architectures either require additional sensor modalities or focus on short-horizon behaviors rather than jointly modeling long-term planning and immediate control.

In reinforcement learning, the Decision Transformer [9] recasts control as sequence modeling conditioned on desired returns. Building on this, the Planning Transformer [1] introduces *Planning Tokens*: coarse latent tokens predicted at regular, long-horizon intervals to capture high-level intent and guide subsequent low-level policies, thereby mitigating compounding error in autoregressive control.

This proposal will unify these streams by integrating Planning Tokens into a vision-only Transformer for driving. We will predict dual time-scale tokens—long-horizon intent tokens interleaved with frame and state tokens—to jointly perform trajectory planning and control within one end-to-end model, combining the strengths of implicit planning with reactive control in complex traffic scenarios.

Project Definition and Objectives

The goal of this project is to develop a vision-only Transformer that unifies trajectory planning and control for autonomous vehicles by integrating long-horizon Planning Tokens [1]. Unlike modular pipelines, our model will learn implicit planning and reactive control in a single end-to-end network.

Specifically, we aim to achieve the following.

- **Architecture Design:** Extend a multi-head self-attention Transformer to process sequences of image frames and vehicle state, predicting Planning Tokens at coarse intervals

alongside low-level control outputs.

- **Implementation:** Build and train the model using the CARLA simulator [2], leveraging its diverse urban scenarios for robust evaluation.
- **Evaluation:** Compare closed-loop driving performance—measured by success rate, trajectory deviation, and collision rate—against a baseline vision-only Transformer without Planning Tokens.
- **Interpretability:** Analyze attention patterns to validate that Planning Tokens attend to long-range road geometry, while control tokens focus on immediate hazards.
- **Reproducibility:** Release code, trained weights, and detailed documentation to facilitate follow-up research.

Methodology

Our expected approach is given as follows.

- **Data Collection & Preprocessing:** We use the CARLA simulator [2] to record sequences of monocular frames (at 10 Hz) and vehicle states (speed, heading). Frames are resized to 128×128 and normalized; state vectors are linearly projected to the same embedding dimension.
- **Tokenization & Embedding:** Each frame and state at time t becomes a token embedding with added sinusoidal positional encodings. At fixed intervals (K timesteps), we insert learnable *Planning Tokens* as in [1] to capture coarse, long-horizon intent.
- **Transformer Backbone:** We stack L layers of multi-head self-attention and feed-forward sublayers [10], operating on the mixed sequence of frame, state, and Planning Tokens.
  - *Planning-Token Head*: MLP predicting latent intent vectors at each Planning Token position.
  - *Control-Token Head*: MLP producing steering, throttle, and brake commands at each regular token.
- **Training Regime:** We perform imitation learning on expert demonstrations, minimizing a combined loss:

$$\mathcal{L} = \lambda_{\text{ctrl}} \, \text{MSE}(\hat{u}, u^*) \; + \; \lambda_{\text{plan}} \, \text{MSE}(\hat{p}, p^*)$$

  where $\hat{u}$ are predicted controls, u^* ground-truth controls, $\hat{p}$ predicted Planning Tokens, and p^* coarse future waypoints clustered every K steps. Optimization uses AdamW with cosine-annealed learning rate.
- **Evaluation:** In closed-loop CARLA scenarios, we measure success rate, trajectory deviation, and collision rate. We compare against a baseline vision-only Transformer without Planning Tokens to quantify the impact of hierarchical intent guidance.

Evaluation Plan

We will rigorously assess our Vision-Only Planning Transformer in closed-loop driving tasks using the CARLA simulator [2], focusing on how Planning Tokens improve long-horizon behavior and robustness.

1. **Closed-Loop Benchmarks:**

   - *Routes:* Three predefined urban routes of increasing length (1 km, 2 km, 5 km), incorporating intersections, turns, and traffic lights.
   - *Traffic Density:* Low (5 vehicles), medium (15 vehicles), and high (30 vehicles) randomized traffic agents.

   We record *success rate* (route completion), *trajectory deviation* (mean lateral and longitudinal error), and *collision rate*.

2. **Baseline Comparison:** Compare against a identical vision-only Transformer without Planning Tokens to quantify the benefit of hierarchical intent guidance.

3. **Ablation Studies:** Vary Planning Token interval $K \in \{5, 10, 20\}$ to study the trade-off between long-horizon guidance and model complexity, measuring performance metrics above and compounding error reduction as in [1].

4. **Attention Visualization:** Generate attention maps to verify that:

   - Planning Tokens attend to distant landmarks (e.g., upcoming intersections).
   - Control tokens attend to nearby obstacles and lane markings.

5. **Statistical Significance:** For each metric, run 30 trials per scenario and perform paired t-tests ($\alpha = 0.05$) to validate significant improvements when using Planning Tokens.

Implementation Feasibility and Timeline

The proposed work leverages existing open-source tools (CARLA, PyTorch) and UTD's HPC cluster to ensure rapid development and reproducibility. Below is a six-month plan:

Months	Tasks
1–2	• Set up CARLA data pipelines; collect and preprocess image/state sequences. • Implement tokenization with interleaved Planning Tokens.
3–4	• Build Transformer backbone and dual heads. • Train baseline and Planning-Token model.
5	• Conduct closed-loop evaluations, ablations on Planning Token interval. • Generate attention visualizations.
6	• Statistical analysis and significance testing. • Documentation, code release, and thesis writing.

Feasibility:

- *Compute:* UTD's Juno cluster provides ample GPU capacity for both training and evaluation.
- *Software:* CARLA and PyTorch ecosystems support rapid iteration and simulation-in-the-loop testing.
- *Expertise:* The IRVL research team's prior experience with learning perception and robotic simulation ensures timely progress.

Expected Outcomes and Deliverables

Upon project completion, we will deliver:

- **Code Repository:** Well-documented PyTorch implementation of the Vision-Only Planning Transformer with Planning Tokens and baseline model.
- **Trained Models:** Checkpoints for both the Planning-Token and non–Planning-Token variants, including scripts for inference in CARLA.
- **Evaluation Report:** Detailed metrics (success rates, trajectory deviation, collision rates) across scenarios, ablation results on token interval, and statistical significance analyses.
- **Attention Visualizations:** Heatmaps demonstrating high-level Planning Tokens attending to long-range landmarks and low-level control tokens focusing on immediate hazards.
- **Thesis Document & Presentation:** Concise write-up of methodology, results, and insights; slides for defense and a possible conference submission.

Conclusion

This proposal outlines the development of a Vision-Only Planning Transformer that will unify long-horizon trajectory planning and low-level control in a single end-to-end model. By interleaving learnable Planning Tokens with frame and state embeddings, we expect to mitigate compounding error and achieve superior route completion, reduced trajectory deviation, and lower collision rates in dense urban scenarios compared to a baseline without hierarchical intent guidance. We will validate this through closed-loop CARLA simulations and statistical analyses, and anticipate that attention visualizations will reveal Planning Tokens attending to distant landmarks while control tokens focus on immediate hazards. Successful completion of this work will demonstrate the practicality of implicit hierarchical planning in vision-only architectures and pave the way for extensions to multi-camera setups and real-world vehicle deployment.

References

[1] Joseph Clinton and Robert Lieck. Planning transformer: Long-horizon offline reinforcement learning with planning tokens, 2024.

[2] Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. CARLA: an open urban driving simulator. *CoRR*, abs/1711.03938, 2017.

[3] Brian Paden, Michal Cáp, Sze Zheng Yong, Dmitry S. Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *CoRR*, abs/1604.07446, 2016.

[4] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual review of control, robotics, and autonomous systems*, 1(1):187–210, 2018.

[5] Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988.

[6] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.

[7] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving, 2022.

[8] Mayank Bansal, Alex Krizhevsky, and Abhijit S. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *CoRR*, abs/1812.03079, 2018.

[9] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *CoRR*, abs/2106.01345, 2021.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.