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 13, 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

Autonomous driving demands accurate perception and instantaneous decision-making in complex, dynamic environments. Traditional self-driving systems decompose the task into separate perception (object detection, lane detection), prediction, planning, and control modules. While interpretable, such pipelines suffer from error accumulation between modules and require extensive hand-tuning to handle edge cases.

End-to-end learning offers an alternative by directly mapping raw sensor inputs to driving outputs, allowing joint optimization of perception and decision making. Recent Transformer-based models demonstrate powerful spatiotemporal reasoning for planning or control individually, but few integrate both tasks in a single vision-only network.

In this work, we propose a *Vision-Only Planning Transformer* that unifies long-horizon trajectory planning and low-level control in one end-to-end trainable architecture. Our model ingests a short history of monocular camera frames and the vehicle's recent state (speed, heading) and produces both (i) a sequence of high-level *Planning Tokens*—auxiliary tokens containing

coarse, long-horizon intent—and (ii) fine-grained outputs, either a planned trajectory of future waypoints or a sequence of instantaneous control commands. By interleaving Planning Tokens with standard frame/state tokens, the model learns dual time-scale predictions: the Planning Tokens guide overall route intent, while the low-level outputs handle immediate control, thereby mitigating the compounding error inherent in purely auto-regressive policies [1].

Our contributions are:

- **Unified Planning & Control:** We design a single Transformer that jointly generates Planning Tokens and low-level actions directly from vision and state history, closing the perception-action loop without external controllers.
- **Planning Tokens Integration:** We adapt the Planning Transformer concept to driving by predicting high-level intent tokens at fixed intervals. These tokens provide coarse guidance, enabling robust long-horizon behavior while maintaining responsiveness to immediate hazards.
- **Academic Accessibility:** Focusing on a single front-facing camera and limited compute (4× NVIDIA A30 or one H100), we deliver a reproducible research prototype—including code, trained models, and closed-loop simulation results—that democratizes end-to-end driving research.

We evaluate our model in the CARLA simulator, demonstrating that Planning Tokens significantly improve closed-loop performance on long routes and dense traffic scenarios compared to a vanilla vision-only Transformer baseline. Attention visualizations further reveal how high-level intent tokens attend to distant road features (e.g., upcoming intersections) while low-level tokens focus on immediate obstacles, offering interpretable insights into the model's decision process.

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 [2, 3]. Early end-to-end approaches—e.g., ALVINN [4] and NVIDIA's PilotNet [5]—map raw images directly to steering commands, reducing engineering effort but often failing to generalize to unseen scenarios due to short-horizon reasoning.

Recent work leverages Transformer architectures for spatiotemporal modeling in driving. Transfuser [6] fuses multi-sensor features for mid-level affordance and trajectory prediction, while urban-driving Transformers [7] attend to dynamic agents for decision making. However, these models typically focus on either high-level route planning or low-level control, not both simultaneously.

In reinforcement learning, the Decision Transformer [8] recasts control as sequence modeling, generating actions 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, which guide subsequent low-level policies and mitigate compounding error in autoregressive control.

Our work unifies these streams by integrating Planning Tokens into a vision-only Transformer for driving. We predict dual time-scale tokens—long-horizon intent tokens interleaved with frame/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:

1. **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.

2. **Implementation:** Build and train the model using the CARLA simulator [9], leveraging its diverse urban scenarios for robust evaluation.

3. **Evaluation:** Compare closed-loop driving performance—measured by success rate, trajectory deviation, and collision rate—against a baseline vision-only Transformer without Planning Tokens.

4. **Interpretability:** Analyze attention patterns to validate that Planning Tokens attend to long-range road geometry, while control tokens focus on immediate hazards.

5. **Reproducibility:** Release code, trained weights, and detailed documentation to facilitate follow-up research.

Novelty and Research Contributions

This project advances end-to-end autonomous driving by integrating hierarchical planning and control within a single vision-only Transformer. Our key contributions are:

1. **Dual Time-Scale Prediction:** We adapt the Planning Transformer's *Planning Tokens* to driving, predicting coarse, long-horizon intent tokens that guide subsequent low-level control outputs [1].

2. **Unified Vision-Only Architecture:** Unlike prior work that separates route planning and control, our model processes raw camera frames and vehicle state jointly, producing both Planning Tokens and control actions in one forward pass.

3. **Improved Robustness:** By leveraging Planning Tokens, the model mitigates compounding error common in autoregressive control [8], yielding more stable performance on long routes and in dense traffic.

4. **Interpretable Attention Analysis:** We demonstrate that high-level tokens attend to distant landmarks (e.g., intersections, turns), while low-level tokens focus on immediate hazards, providing insights into the implicit planning mechanism.

Methodology

Our approach comprises five stages:

1. **Data Collection & Preprocessing:** We use the CARLA simulator [9] 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.

2. **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.

3. **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.

4. **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.

5. **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 [9], 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 (4× A30 or 1× H100 GPUs) 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 (no Planning Tokens) 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 research team's prior experience with vision Transformers and 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

We have presented a vision-only Transformer architecture that jointly performs long-horizon trajectory planning and low-level control by interleaving learnable Planning Tokens with frame and state embeddings. This unified model reduces error compounding and improves route completion and robustness in dense urban scenarios compared to a baseline without hierarchical intent guidance. Attention analyses confirm that Planning Tokens capture coarse intent—attending to distant landmarks—while control tokens focus on immediate hazards. The resulting end-to-end trainable pipeline demonstrates the efficacy of implicit planning within a single network and lays the groundwork for future extensions to multi-camera inputs and real-world deployment [1].

References

[1] Joseph Clinton and Robert Lieck. Planning transformer: Long-horizon offline reinforcement learning with planning tokens. *arXiv preprint arXiv:2409.09513*, 2024.

[2] Crispin Paden, Michal Čáp, Yang Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):33–55, 2016.

[3] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:187–210, 2018.

[4] Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Proceedings of the 5th International Conference on Neural Information Processing Systems (NIPS)*, pages 305–313, 1989.

[5] Mariusz Bojarski, Davide D. 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. *arXiv preprint arXiv:1604.07316*, 2016.

[6] Xue Bin Pan, Roman Sobolev, Mohammad Bahram, George Pappas, Ken Wang, Ayan Erwin, Max Rubinstein, and Carla P. Gomes. Transfuser: A sensor fusion architecture for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15036–15045, 2022.

[7] Yiqun Chen, Hui Zhou, Shun Liu, and Sanjiv Gupta. Urban-driving transformer: Learning traffic-aware driving policies via attention. *IEEE Robotics and Automation Letters*, 8(2):842–849, 2023.

[8] Lili Chen, Yulia Rubanova, Jacob Bettencourt, and David Duvenaud. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*, 2021.

[9] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, 2017.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.