CAUSALDRIVE: BREAKING MOTION INERTIA IN VLA-BASED AUTONOMOUS DRIVING BY ALIGNED SEMANTIC INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As End-to-End autonomous driving advances into the Vision-Language-Action (VLA) era, models increasingly rely on natural language to generate human-aligned control. Despite this progress, a critical vulnerability remains: *inertial driving*, defined as the tendency to maintain historical motion patterns even when facing sudden hazards. We identify three factors underlying this issue: *memory bottlenecks in temporal fusion*, *inertial bias in implicit temporal modeling*, and *decoupling of reasoning and action*, which together allow misleading temporal inertia to dominate decision-making. To address this issue, we introduce **CausalDrive**, a novel VLA framework that explicitly bridges semantic reasoning and temporal action generation through language-guided causal intervention. Concretely, our architecture features three core innovations: 1) *Streaming Sparse Temporal Fusion (SSTF)* for maintaining efficient, long-term context via lightweight memory slots; 2) *Task-Query-Guided Temporal Modeling* to generate kinematically consistent ego-trajectories; and 3) *Causal-Intervention Attention (CIA)*, a mechanism that dynamically overrides historical momentum when triggered by safety-critical language cues (*e.g.*, "ignore momentum"). We further optimize CausalDrive through supervised fine-tuning and constrained Group Relative Policy Optimization (GRPO) guided by learned Reward/Cost Models. Extensive experiments demonstrate its effectiveness, achieving 87.41% accuracy on DriveAction, 0.30m trajectory error on DriveLM, and 0.0142 steering MAE on BDDA. Qualitative results further show its ability to suppress inertial bias during critical events, paving the way for highly reliable and interpretable autonomous driving systems.

1 INTRODUCTION

End-to-end (E2E) autonomous driving seeks to map sensory observations directly to driving decisions, reducing the dependence on manually designed interfaces between perception, prediction, planning, and control Chen et al. (2024); Dong et al. (2025). Earlier systems mainly focused on direct control prediction or compact intermediate driving representations Bojarski et al. (2016); Hu et al. (2023b). As driving scenarios became more complex, these formulations showed limited capability in high-level semantic reasoning. More recent work therefore began to integrate large vision-language models (VLMs) into autonomous driving, leading to Vision-Language-Action (VLA) systems that use language to interpret complex scenarios and generate human-aligned behaviors Yang et al. (2023); Jiang et al. (2023). Despite their stronger reasoning capability, current VLA models can still exhibit a critical vulnerability: *inertial driving*, where models fail to override historical motion patterns when facing sudden hazards (*e.g.*, a pedestrian suddenly entering the lane in Figure 1), leading to catastrophic failures despite correct linguistic reasoning.

This failure does not stem from a single component, but reflects several limitations across the current VLA driving pipeline. We identify three factors that are particularly relevant to inertial behavior: 1) **Memory Bottlenecks in Temporal Fusion:** Traditional approaches rely on *dense temporal fusion* (*e.g.*, stacking multi-frame BEV features), which suffers from quadratic memory growth and error accumulation, severely limiting the horizon for long-range

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

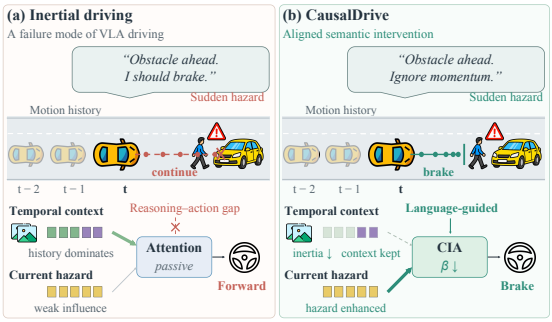


Figure 1: Illustration of inertial driving and CausalDrive. (a) Inertial driving: historical motion dominates action generation despite correct hazard recognition. (b) CausalDrive: language-guided intervention suppresses misleading motion inertia and emphasizes safety-critical current cues.

reasoning Wang et al. (2022). While recent sparse streaming architectures improve efficiency Jia et al. (2025), they merely propagate task queries without addressing the core behavioral flaw: the inability to distinguish and discard misleading history when the driving context shifts. 2) **Inertial Driving in Implicit Recurrent Modeling:** Methods utilizing RNNs or latent world models Zhang et al. (2026); Li et al. (2025) often conflate causally relevant present cues with historical momentum. This leads to a pervasive failure mode where the agent continues along its prior trajectory even when confronted with emergent hazards (*e.g.*, jaywalking pedestrians), as the hidden states inherently reinforce temporal continuity over safety-critical discontinuity. 3) **Decoupling of Reasoning and Action:** Recent reasoning-augmented models generate post-hoc Chain-of-Thought (CoT) explanations NVIDIA et al. (2026); Yuan et al. (2025), yet their linguistic reasoning remains structurally decoupled from the action-generation attention mechanism. Consequently, a model may output "brake for pedestrian" while its trajectory decoder still acts on historical velocity, resulting in delayed responses. This underscores a critical gap: *without explicit neural mechanisms to suppress inertia, causal awareness alone is insufficient for robust control*.

To address these interconnected challenges, we propose **CausalDrive**, a novel VLA framework that redefines temporal modeling through sparsity, query evolution, and causal intervention. Our approach introduces three core innovations: First, to alleviate memory bottlenecks, we devise **Streaming Sparse Temporal Fusion (SSTF)**. This module replaces dense BEV stacking with a lightweight *Semantic Memory Slot* that retains compact semantic representations from the final layer of a frozen VLM encoder. This enables efficient, low-latency inference while supporting heterogeneous inputs ranging from short commands to long video sequences. Second, to overcome recurrent instability, we introduce **Task-Query-Guided Temporal Modeling**. By framing driving as the continuous evolution of a learnable *Ego Query*, our model implicitly aligns motion primitives across time via self-attention, ensuring kinematically consistent trajectories without the need for unstable RNNs or ad-hoc smoothing heuristics. Most critically, to bridge the disconnect between reasoning and action, we design the **Causal-Intervention Attention (CIA)**. Guided by natural language causal assertions (*e.g.*, "ignore momentum"), CIA parses discrete state-change signals to dynamically generate a modulation factor β. This factor actively suppresses momentum-related features in historical queries while amplifying attention to salient current-frame tokens. In effect, CIA achieves *adaptive forgetting*, retaining useful context while discarding misleading inertia to enable rapid, physically grounded reactions in emergencies. Furthermore, to align the learned policy with safety-critical decision-making, we introduce a two-stage optimization strategy. The model is first initialized through Supervised Fine-Tuning (SFT) using expert demonstrations, and then refined through constrained Group Relative Policy Optimization (GRPO) Shao et al. (2024) with learned Reward/Cost Models. This alignment process enables the policy to balance task performance and safety constraints while further improving its ability to respond to critical state changes.

Extensive experiments on DriveAction Hao et al. (2025), DriveLM Sima et al. (2024), and BDDA Yu et al. (2020) demonstrate that CausalDrive significantly outperforms prior methods in both standard planning metrics and safety-critical edge cases. Further qualitative and quantitative analysis confirms that explicit sparsity, query-driven dynamics, and causal intervention are essential for building truly reliable end-to-end autonomous systems.

In summary, our main contributions are threefold:

- **A Unified Sparse Streaming Architecture:** We propose SSTF, replacing computationally expensive dense temporal stacking with a lightweight Semantic Memory Slot, enabling efficient long-term context retention and seamless support for heterogeneous input modalities.

- **Kinematically Consistent Query Evolution:** We introduce Task-Query-Guided Temporal Modeling, which formulates trajectory prediction as the evolution of a learnable Ego Query, implicitly aligning historical states via self-attention to ensure smooth and physically plausible motion primitives.
- **Causal Intervention for Safety-Critical Control:** We design the Causal-Intervention Attention (CIA) module to explicitly bridge linguistic reasoning and action execution. By dynamically modulating attention weights based on causal assertions, CIA empowers the model to actively suppress misleading historical inertia, significantly reducing trajectory errors in emergency scenarios.

2 Related Work

2.1 Temporal Modeling and the Passive Memory Bottleneck

Early End-to-End (E2E) driving systems primarily relied on dense temporal fusion, concatenating past frames along channel or time dimensions, or accumulating Bird's-Eye-View (BEV) features Zarghani et al. (2025); Chen et al. (2015); Dosovitskiy et al. (2017); He et al. (2016); Li et al. (2022). While capturing rich spatio-temporal correlations, these dense methods suffer from quadratic memory growth and error propagation, limiting long-horizon reasoning Wang et al. (2022); Jia et al. (2025). To alleviate this computational bottleneck, subsequent paradigms shifted toward implicit history compression and sparse streaming architectures. Recurrent Neural Networks (RNNs) Zhang et al. (2025); Mambelli et al. (2024); Lv et al. (2025); Zhang et al. (2026) and generative latent world models Li et al. (2025); Hu et al. (2023a); Wang et al. (2023b) encode temporal dynamics into hidden states or future scene forecasts. Concurrently, sparse architectures like StreamPETR Wang et al. (2023a), DriveTransformer Jia et al. (2025), and OmniReason Liu et al. (2025) achieve efficient long-term context by propagating learnable task queries across frames.

Despite their computational efficiency and elegance, both implicit and sparse methods share a fundamental vulnerability: they operate under a passive memory paradigm. Because they retain and process historical information uniformly, they reinforce the model's reliance on past momentum. Consequently, when present contexts contradict historical trends—such as the sudden appearance of an obstacle—these models struggle to adapt promptly, manifesting as the *inertial driving* behavior that fundamentally plagues current E2E systems.

2.2 Causal Reasoning and Intervention in VLAs

The rise of Vision-Language-Action (VLA) models has brought causal reasoning and interpretability to the forefront of autonomous driving. Models like Alpamayo-R1 NVIDIA et al. (2026) introduce Chain-of-Causation (CoC) traces to ground decisions, while AutoDrive-R^2 Yuan et al. (2025) enforces self-reflection through physics-aware reinforcement learning. Furthermore, systems such as OmniReason Liu et al. (2025) and recent explainable frameworks Wang et al. (2024) generate natural language rationales (*e.g.*, "Stop at red light") to justify actions and enhance user trust.

While these works represent a significant leap toward human-aligned autonomy, a crucial gap remains: the generated reasoning is often structurally decoupled from the action-generation mechanism. A linguistic module may correctly identify a hazard, yet the underlying attention weights in the trajectory decoder remain dominated by historical velocity features. Without an explicit neural intervention mechanism to modulate these weights based on causal language assertions, the model cannot guarantee that its physical actions align with its high-level reasoning. Resolving this critical disconnect between cognition and control is the core motivation behind our proposed architecture.

3 Method

Preliminaries. End-to-end autonomous driving aims to predict the future trajectory of the ego vehicle directly from onboard visual observations, optionally conditioned on a natural language instruction. At time step t, the model receives the visual observations available up to the current time, $\{I_\tau\}_{\tau=1}^{t}$, together with an optional instruction c (*e.g.*, "Turn left at the intersection"), and predicts a future ego trajectory $\hat{W}_t = \{\hat{w}_t^1, \ldots, \hat{w}_t^H\}$ over a prediction horizon of H steps Tian et al. (2024). Each

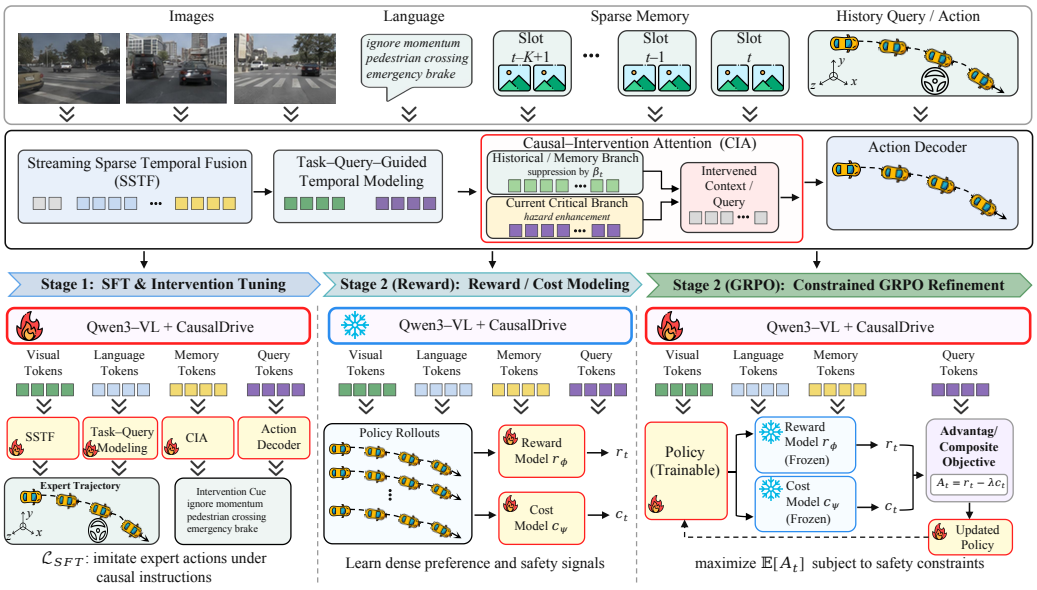


Figure 2: **Overview of CausalDrive architecture and training pipeline.** (Left) Inference Flow: multi-view visual observations and language instructions are encoded into sparse temporal representations through SSTF and Task-Query-Guided Temporal Modeling. CIA further performs language-guided intervention by suppressing misleading historical motion features and enhancing safety-critical current cues before trajectory decoding. (Right) Training Pipeline: the full CausalDrive architecture is first initialized with expert demonstrations through SFT, followed by Reward/Cost Modeling and constrained GRPO to align the policy with task performance and safety constraints.

waypoint $\hat{w}_t^h \in \mathbb{R}^2$ denotes the predicted Bird's-Eye-View (BEV) position of the ego vehicle at time step $t + h$.

Effective trajectory prediction requires exploiting temporal context while remaining responsive to newly observed events that may invalidate the previous motion tendency. We therefore propose CausalDrive, a VLA framework designed to model temporal context while explicitly reducing the influence of outdated motion information when the driving state changes.

3.1 Architecture

Overview. As illustrated in Figure 2, CausalDrive follows an encoder–decoder architecture for language-conditioned trajectory prediction. The current multi-view observation and optional language instruction are first encoded by a frozen Qwen3-VL backbone Bai et al. (2025) to obtain visual-language features. Temporal information is incorporated through SSTF, which maintains semantic context from recent observations, and Task-Query-Guided Temporal Modeling, which captures the evolution of the ego-driving state across time. CIA further integrates the temporal context with the current visual-language features and adaptively regulates the contribution of historical motion information according to the current driving context. The fused representation is finally decoded into the future ego trajectory $\hat{W}_t$.

3.1.1 Streaming Sparse Temporal Fusion (SSTF)

Traditional approaches often rely on *dense temporal fusion*, stacking multi-frame BEV features or concatenating raw images over time Li et al. (2024); Huang & Huang (2022). While expressive, this strategy incurs *quadratic memory growth* ($O(T^2)$) and accumulates encoding errors over long horizons, making real-time inference on embedded devices impractical. Conversely, simply truncating history leads to a loss of critical long-term context.

To resolve this dilemma, we introduce a *Semantic Memory Slot* mechanism that compresses historical information without losing semantic fidelity. Let f_{VLM} denote the frozen Qwen3-VL Bai et al. (2025)

encoder. Given the current frame I_t at time step t and a textual command c, we extract

$$z_t = \mathrm{FinalLayer}\big(f_{\mathrm{VLM}}(I_t, c)\big) \in \mathbb{R}^{N\times d}, \tag{1}$$

where $\mathrm{FinalLayer}$ extracts the hidden representations from the final Transformer layer, N denotes the number of multimodal tokens, and d represents the hidden dimension. Instead of retaining dense historical features, we store these compact semantic representations in a fixed-size first-in-first-out (FIFO) memory queue with capacity K, enabling efficient streaming temporal context aggregation:

$$\mathcal{M}_t = \begin{cases} \mathcal{M}_{t-1} \parallel z_t, & |\mathcal{M}_{t-1}| < K, \\ \mathrm{Tail}(\mathcal{M}_{t-1}) \parallel z_t, & |\mathcal{M}_{t-1}| = K, \end{cases} \qquad \mathcal{M}_0 = \emptyset, \tag{2}$$

here, $|\mathcal{M}_t|$ denotes the number of stored representations, $\parallel$ denotes appending a representation to the end of the queue, and $\mathrm{Tail}(\cdot)$ removes the oldest entry from the FIFO memory.

This design reduces memory complexity to linear $O(K)$, enabling low-latency inference even with long context windows. Furthermore, the fixed-size sliding window naturally supports heterogeneous input modalities, seamlessly handling both single-image commands (DriveAction) Hao et al. (2025) and long video sequences (BDDA) Yu et al. (2020) without architectural modifications.

3.1.2 Task-Query-Guided Temporal Modeling

Modeling temporal dynamics in driving is challenging. Recurrent Neural Networks (RNNs) and 3D CNNs, commonly used in prior works, often suffer from *gradient vanishing/explosion* over long sequences and lack global coherence, leading to jittery or physically implausible trajectories Jia et al. (2025); Zhang et al. (2026). Without a mechanism to explicitly align motion intentions across time, the model may generate erratic control signals that violate vehicle kinematics.

To overcome these instabilities, we reframe trajectory prediction as the continuous evolution of a learnable *Ego Query* $q_t^{\text{ego}} \in \mathbb{R}^{H\times d}$, where H denotes the number of future prediction steps and d is the hidden dimension. This query explicitly encodes the vehicle's current spatio-temporal state and dynamical intent. Instead of relying on hidden states passed sequentially, the current query interacts with its historical instances $\mathcal{Q}^{\text{hist}} = [q_{t-K+1}^{\text{ego}}, \ldots, q_{t-1}^{\text{ego}}]$ via a global self-attention mechanism, where K denotes the temporal history length:

$$\tilde{q}_t^{\text{ego}} = \mathrm{SelfAttn}\left(q_t^{\text{ego}}, \mathcal{Q}^{\text{hist}}, \mathcal{Q}^{\text{hist}}\right), \tag{3}$$

where, q_t^{ego} serves as the current query, $\mathcal{Q}^{\text{hist}}$ provides the historical keys and values, and $\tilde{q}_t^{\text{ego}}$ denotes the temporally enhanced ego-query representation.

This global interaction allows the model to implicitly align motion primitives across all time steps simultaneously, rather than step-by-step. The result is a set of trajectories that are not only smooth but also *kinematically consistent* over the prediction horizon $t \to t+H$, providing a robust alternative to unstable recurrent dynamics or ad-hoc smoothing heuristics.

3.1.3 Causal-Intervention Attention (CIA)

A critical failure mode in current VLAs is *inertial driving*: even when the language module correctly identifies a hazard (*e.g.*, "Brake!"), the underlying attention mechanism remains dominated by historical velocity features, causing the vehicle to continue moving forward. This disconnect arises because standard attention passively aggregates history; it lacks an explicit mechanism to *override* misleading momentum when safety is at stake.

To bridge this gap, we introduce a neural intervention mechanism that dynamically modulates attention weights based on causal assertions. A dedicated head generates explanations (*e.g.*, "Pedestrian ahead, ignore momentum"), from which we extract state mutation signals to compute a modulation factor $\beta_t \in (0, 1]$:

$$\beta_t = \sigma\left(w^\top \cdot \mathrm{avgpool}(z_t^{\text{critical}}) + b\right), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, and z_t^{critical} represents embeddings of detected hazard tokens. When a hazard is detected, β_t approaches 0, triggering an active suppression of momentum features in the historical value matrix:

$$V^{\text{mod}} = V \odot \left(1 - (1-\beta_t)\cdot \mathbb{I}_{\text{mom}}\right). \tag{5}$$

Here, $\mathbb{I}_{\text{mom}}$ masks tokens associated with inertial motion. Simultaneously, we amplify attention to current hazards by modifying the keys:

$$K^{\text{mod}} = K + \gamma \cdot \mathbb{I}_{\text{critical}}, \tag{6}$$

where $\gamma > 0$ boosts the attention scores for emergent obstacles.

This dual-modulation strategy realizes *adaptive forgetting*: the model retains useful historical context during normal driving but instantly discards misleading inertia when safety costs spike. This ensures that linguistic reasoning directly translates into physical action, enabling rapid, life-saving reactions in emergency scenarios.

3.2 Training Pipeline

Training a safe and reliable VLA requires more than simple imitation. It demands explicit optimization of safety objectives that are difficult to capture through supervised learning alone (Achiam et al., 2017; Shalev-Shwartz et al., 2017; Yuan et al., 2025). Therefore, CausalDrive is optimized through a progressive two-stage pipeline: 1) **Supervised Fine-Tuning (SFT)** to establish a competent behavioral baseline, and 2) **Constrained GRPO with Learned Reward/Cost Models** to refine the policy towards safety and preference alignment. This approach avoids the instability of training RL from scratch while overcoming the rigidity of rule-based safety filters.

3.2.1 Stage 1: Supervised Fine-Tuning

SFT provides a crucial warm-start by learning from expert demonstrations, ensuring that the policy initially exhibits expert-aligned driving behaviors before safety-oriented policy optimization (Bojarski et al., 2016; Shao et al., 2024). This initialization avoids optimizing an under-trained policy and provides a stable foundation for subsequent reward-driven refinement. The stage is particularly important for CausalDrive, as the proposed temporal modeling and intervention mechanisms require meaningful driving representations before further safety-aware refinement.

We initialize the policy π_θ by minimizing the deviation from expert demonstration data $\mathcal{D}_{\text{expert}}$:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(I,c,W^*)\sim\mathcal{D}_{\text{expert}}} \left[\|\hat{W}_t - W^*\|_1 \right], \tag{7}$$

where W^* denotes the ground-truth trajectory. Importantly, the full CausalDrive architecture, including SSTF, Ego Query, and CIA (*cf.* §3.1), is optimized during this stage. Therefore, SFT not only learns the action prediction objective but also initializes the sparse temporal memory, query-based motion representation, and language-guided intervention mechanism from expert driving behaviors. This provides a reliable starting point for subsequent constrained GRPO optimization, where the policy is further refined under explicit safety constraints.

3.2.2 Stage 2: Constrained GRPO with Learned Reward/Cost Models

Standard SFT primarily learns from imitates expert demonstrations and cannot explicitly optimize safety preferences in out-of-distribution scenarios. While some works use hard-coded geometric rules for safety, these are often too rigid and fail to capture nuanced driving preferences Shalev-Shwartz et al. (2017); Lu et al. (2025). We instead adopt a *learned constraint* approach within the Group Relative Policy Optimization (GRPO) Shao et al. (2024) framework, allowing the model to discover optimal trade-offs between efficiency and safety.

Learning Flexible Reward and Cost Functions. Before optimizing the policy, we attach two lightweight heads to the SFT-initialized model to learn dense feedback signals from data:

- **Reward head** r_ϕ: Predicts task reward r_t (*e.g.*, route progress, comfort), capturing human preferences that are hard to specify manually.
- **Cost head** c_ψ: Predicts safety cost c_t, defined as:

$$c_t = \mathbb{I}_{\text{collision}} + \alpha \cdot \mathbb{I}_{\text{traffic_violation}} + \beta \cdot \text{ttc_penalty}, \tag{8}$$

  where $\mathbb{I}$ denotes indicator functions for discrete violations, and α, β weight continuous risks like Time-To-Collision (TTC).

Both heads are trained via regression. Notably, the **Causal Assertion Head** (from CIA) generates natural language explanations that supervise the cost head, grounding abstract cost values in interpretable semantic events (*e.g.*, "near miss with pedestrian").

Policy Optimization with Safety Constraints. While SFT provides a competent behavioral initialization, it mainly captures the average behavior from expert demonstrations and cannot explicitly optimize safety preferences. To further align the policy with safety-critical decision-making, we refine π_θ using GRPO Shao et al. (2024) under learned reward and cost constraints. GRPO is adopted to optimize the policy through relative advantages without introducing an additional critic network. Specifically, the objective is formulated as:

$$\max_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \quad \text{s.t.} \quad \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c_t \right] \leq d. \tag{9}$$

We solve this constrained optimization problem using Lagrangian relaxation Achiam et al. (2017), resulting in the following objective:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t (r_t - \lambda c_t) \right] + \lambda d, \tag{10}$$

where $\lambda \geq 0$ controls the trade-off between task performance and safety. We dynamically update λ through dual ascent:

$$\lambda \leftarrow \lambda + \eta_\lambda \left(\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c_t \right] - d \right). \tag{11}$$

This formulation allows λ to automatically adjust the penalty strength: if the policy violates safety constraints, λ increases, forcing the model to prioritize safety over reward. During this process, our architectural innovations are fully leveraged: SSTF ensures efficient rollouts, Ego Query maintains trajectory smoothness, and CIA actively intervenes to reduce the cost signal c_t by suppressing inertia in critical moments.

4 Experiments

4.1 Experimental Setup and Training Philosophy

Benchmarks. To verify the robustness and generalization capability of CausalDrive, we evaluate it on three widely-utilized benchmarks, each focusing on a distinct aspect of the VLA pipeline:

- **DriveAction** Hao et al. (2025): Focuses on high-level semantic action understanding. The model is provided with a **single scene image** and a **textual instruction** (*e.g.*, "Turn left at the next signal"). The task is to categorize the visual cues into one of the discrete action classes, and the metric is Action Accuracy (%).
- **DriveLM** Sima et al. (2024): Evaluates mid-to-long-term planning and spatial reasoning. The input consists of **temporal video frames** paired with **natural language reasoning tasks**. The model must predict a future trajectory represented as a sequence of waypoints, and the metric is Average Trajectory L2 Error (meters).
- **BDDA** Yu et al. (2020): Targets low-level vehicle control and dynamic grounding. The input combines **high-frequency video frames** with synchronous **GPS sequences** (odometry and inertial data). The task is to predict continuous control signals such as steering and acceleration, and the metric is Mean Absolute Error (MAE).

Implementation Details. Due to space limitations, details are provided in Appendix §A.

4.2 Quantitative Results

4.2.1 Action Classification and Ablation on DriveAction

Table 1 summarizes the performance on DriveAction Hao et al. (2025), which evaluates the model's ability to understand high-level driving intentions from visual observations and language instructions.

CausalDrive achieves **87.41**% accuracy under the full V-L-A setting, outperforming Claude 3.7 Sonnet and achieving competitive performance with GPT-4o Hurst et al. (2024). This result demonstrates that the proposed framework can effectively integrate visual perception and linguistic reasoning for semantic driving decision-making.

Table 1: DriveAction accuracy (%) and modality ablation.

Model	V-L-A	V-A	L-A
GPT-4o	88.84	84.72	86.52
Claude 3.7 Sonnet	86.31	80.80	82.56
Qwen-Max	91.32	88.38	89.16
CausalDrive (Ours)	**87.41**	**82.17**	**85.35**

We further investigate the contribution of different modalities through ablation. As shown in Table 1, removing language information (V-A setting) decreases the accuracy from 87.41% to 82.17%, highlighting the importance of language grounding for interpreting complex driving scenarios and safety-critical instructions. Meanwhile, the L-A setting achieves 85.35% accuracy, substantially outperforming the V-A setting, indicating that semantic instructions provide strong priors for high-level decision-making, while visual observations offer complementary scene-specific evidence. These results verify the necessity of jointly modeling visual and linguistic cues in VLA-based driving systems.

Training dynamics of the two-stage alignment strategy are further analyzed in Appendix §B.1. The results demonstrate stable optimization of SFT initialization, Reward/Cost Modeling, and constrained GRPO refinement.

4.2.2 Trajectory Planning on DriveLM

Table 2 reports trajectory planning results on DriveLM Sima et al. (2024), which evaluates the model's ability to predict future ego trajectories under visual-language driving scenarios. CausalDrive achieves an average L2 error of **0.30m**, outperforming both general-purpose vision-language models (*e.g.*, LLaMA Grattafiori et al. (2024)) and specialized autonomous driving models. In particular, it reduces the trajectory error from 0.40m and 0.33m of DriveVLM Tian et al. (2024) and OmniDrive Wang et al. (2025); Liu et al. (2025) to 0.30m, respectively, demonstrating more accurate long-horizon trajectory prediction.

Table 2: Trajectory results on DriveLM (Avg. L2 Error ↓).

Model	L2 Error (m)
DeepSeek-VL2-16B	1.75
Qwen-2.5-VL-7B	1.45
LLaMA-3.2-11B-Vision	1.54
DriveVLM	0.40
OmniDrive	0.33
EMMA	0.32
CausalDrive (Ours)	**0.30**

The improvement can be attributed to the explicit temporal modeling and semantic-action alignment in CausalDrive. The Task-Query-Guided Temporal Modeling module maintains kinematically consistent motion evolution, while Causal-Intervention Attention enables the policy to adaptively suppress outdated motion patterns when current driving contexts change. Together, these designs allow the model to generate trajectories that are both temporally coherent and responsive to emerging safety-critical events. Additional optimization dynamics are reported in Appendix §B.1.

4.2.3 Continuous Control on BDDA

Table 3: Low-level control evaluation on BDDA (MAE).

Model	Steering MAE	Accel MAE
FCN-LSTM	$0.025 \sim 0.030$	$0.23 \sim 0.25$
CausalDrive (Ours)	**0.0142**	**0.1908**

On the BDDA Yu et al. (2020) dataset, we evaluate CausalDrive on low-level continuous control, where the model is required to predict fine-grained steering and acceleration commands from sequential driving observations. As shown in Table 3, CausalDrive achieves a Steering MAE of **0.0142** and an Accel MAE of **0.1908**, outperforming the FCN-LSTM baseline Xu et al. (2017) by a large margin.

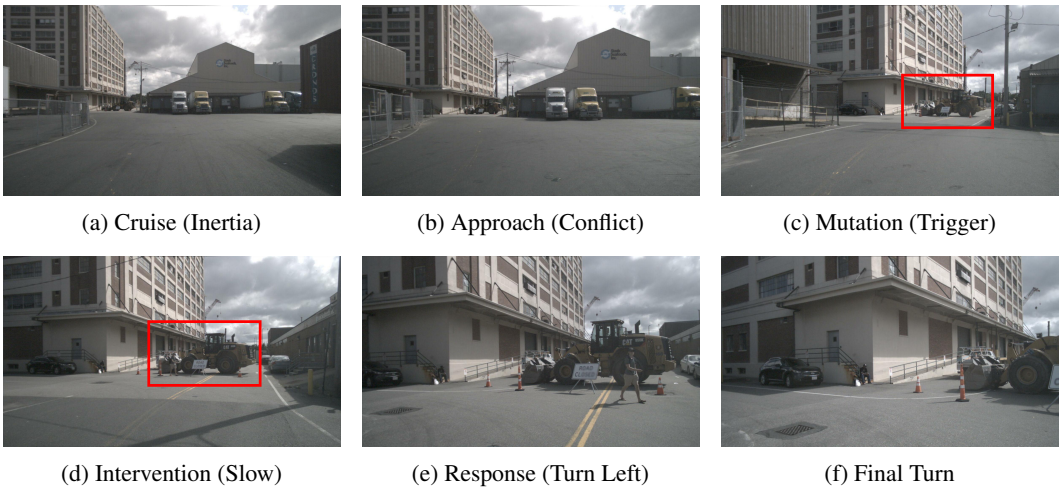
These results demonstrate that the proposed temporal modeling and language-guided intervention mechanism can effectively translate high-level semantic understanding into precise vehicle control.

Beyond imitation accuracy, BDDA further evaluates whether the policy can satisfy explicit safety constraints during optimization. The training dynamics in Appendix §B.1 show that, during the constrained-GRPO stage, the safety constraint fulfillment remains consistently below the predefined 0.15 limit. Meanwhile, the adaptive adjustment of the penalty strength λ enables the policy to dynamically balance control objectives and safety requirements, resulting in stable continuous control under constrained optimization.

4.3 Qualitative Analysis: Causal Intervention Case Study

To highlight our model's unique capability in mitigating the pervasive inertial driving bias found in traditional End-to-End architectures, a scenario where historical momentum directly conflicts with emergent safety constraints is selected. We evaluate the model's ability to override motion inertia through causal reasoning in this challenging "road block" scenario (Figure 3).

Initially, the ego vehicle maintains a steady cruising velocity, *cf.* Steps (a)–(b). At Step (c), a loader suddenly blocks the path, and someone enters the frame. Traditional End-to-End models often suffer from Inertial Bias, tending to continue forward motion based on historical patterns. In contrast, our VLA model identifies these visual elements as causal triggers through linguistic grounding. The GRPO-aligned policy, informed by the Stage 2 cost model, assigns a prohibitive penalty to "continue straight" actions. Consequently, the predicted trajectory collapses to a halt at the current position (Step (d)). This phase transition from cruising to stopping demonstrates that our model performs causal deduction rather than passive imitation, prioritizing safety-critical semantic signals over temporal inertia.



(a) Cruise (Inertia) (b) Approach (Conflict) (c) Mutation (Trigger)

(d) Intervention (Slow) (e) Response (Turn Left) (f) Final Turn

Figure 3: Qualitative sequence of a "continue straight" intervention. Visual grounding of loader and pedestrian at Step (c) triggers a causal intervention, breaking the action inertia by Step (d).

5 Conclusion

This work presents CausalDrive, a Vision-Language-Action framework that explicitly addresses temporal inertia in end-to-end autonomous driving through language-guided causal intervention. By integrating sparse temporal modeling, query-based motion evolution, and semantic intervention into the action-generation process, CausalDrive enables the policy to preserve useful historical context while suppressing misleading motion inertia under safety-critical situations. Experiments across DriveAction, DriveLM, and BDDA demonstrate the effectiveness of CausalDrive in high-level decision-making, trajectory planning, and low-level control. Qualitative analyses further validate that the proposed intervention mechanism allows the model to respond adaptively to abrupt environmental changes, bridging the gap between linguistic reasoning and physical action execution.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. Pmlr, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024.

Y. Chen, H. Hu, and G. Gong. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2722–2730, 2015.

Wei Dong, Sikai Lu, Xinhe Chen, Shunyao Zhang, Qingchao Liu, Ze Liu, Long Chen, Hai Wang, and Yingfeng Cai. End-to-end autonomous driving: From classic paradigm to large model empowerment—a comprehensive survey. *IEEE Internet of Things Journal*, 13(3):3870–3898, 2025.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, pp. 1–16, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, et al. Driveaction: A benchmark for exploring human-like driving decisions in vla models. *arXiv preprint arXiv:2506.05667*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Anthony Hu, Lloyd Russell, Hudson Yeo, et al. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023a.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. pp. 17853–17862, 2023b.

Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. *arXiv preprint arXiv:2503.07656*, 2025.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8350, 2023.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481*, 2025.

Yinhao Li, Zheng Ge, Guanyi Yu, et al. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. 2024.

Pei Liu, Qingtian Ning, Xinyan Lu, Haipeng Liu, Weiliang Ma, Dangen She, Peng Jia, Xianpeng Lang, and Jun Ma. Omnireason: A temporal-guided vision-language-action framework for autonomous driving. *arXiv preprint arXiv:2509.00789*, 2025.

Sikai Lu, Yingfeng Cai, Ze Liu, Yubo Lian, Long Chen, and Hai Wang. A preference-based multi-agent federated reinforcement learning algorithm framework for trustworthy interactive urban autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Lei Lv, Yunfei Li, Yu Luo, et al. Flow-based policy for online reinforcement learning. *arXiv preprint*, 2025.

Davide Mambelli, Stephan Bongers, et al. When do off-policy and on-policy policy gradient methods align? *arXiv preprint*, 2024.

NVIDIA, Yan Wang, Wenjie Luo, Boris Ivanovic, Marco Pavone, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2026.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Shihao Wang, Yingfei Liu, Tiancai Wang, et al. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22442–22452. IEEE, 2025.

Yixuan Wang, Ruochen Jiao, et al. Empowering autonomous driving with large language models: A safety perspective. *arXiv preprint arXiv:2312.00812*, 2024.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning (CoRL)*, pp. 180–191, 2022.

Yuqi Wang, Jiawei He, Lue Fan, et al. Driving into the future: Multiview visual forecasting and planning with world model. *arXiv preprint arXiv:2311.17918*, 2023b.

Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. *arXiv preprint arXiv:1612.01079*, 2017.

Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, and Peng Chen. Bevheight: A robust framework for vision-based roadside 3d object detection. *arXiv preprint arXiv:2303.08498*, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *arXiv preprint arXiv:1805.04687*, 2020.

Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, et al. Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025.

A. Zarghani, A. Ebrahimi, and A. Malekesfandiari. Multimodal framework for explainable autonomous driving. *arXiv preprint arXiv:2507.07938*, 2025.

Tonghe Zhang, Chao Yu, Sichang Su, and Yu Wang. Reinflow: Fine-tuning flow matching policy with online reinforcement learning. *arXiv preprint arXiv:2501.XXXXX*, 2025.

Yixian Zhang, Shu'ang Yu, Tonghe Zhang, Mo Guang, Haojia Hui, Kaiwen Long, Yu Wang, Chao Yu, and Wenbo Ding. Sac flow: Sample-efficient reinforcement learning of flow-based policies via velocity-reparameterized sequential modeling. *arXiv preprint arXiv:2509.25756*, 2026.

APPENDIX

This appendix is organized as follows:

- §A provides additional implementation details, including training configurations, testing protocols, and architectural settings.
- §B.1 presents additional experimental analysis of the two-stage alignment strategy through training dynamics on different benchmarks.

A IMPLEMENTATION DETAILS

Training. Our model is trained on 8 NVIDIA A100 GPUs. We use Qwen3-VL Bai et al. (2025) as the frozen vision-language backbone and optimize the newly introduced CausalDrive modules, including SSTF (§3.1.1), Task-Query-Guided Temporal Modeling (§3.1.2), CIA (§3.1.3), the action decoder, and the reward/cost heads. The model is trained with a global batch size of 32 using bfloat16 mixed precision. We adopt AdamW optimizer with an initial learning rate of 2e-5, weight decay of 0.05, and gradient clipping of 1.0. The training follows a two-stage pipeline: supervised fine-tuning (§3.2.1) is first conducted on expert demonstrations, followed by constrained GRPO (§3.2.2) refinement with learned reward and cost models. By default, the semantic memory length is set to 8 for DriveAction and DriveLM, and 16 for BDDA.

Testing. Testing is conducted on the same hardware as training. No data augmentation is used during evaluation. For DriveAction Hao et al. (2025), the model predicts the discrete driving action with maximum probability. For DriveLM Sima et al. (2024), it predicts a sequence of future BEV waypoints and is evaluated by average trajectory L2 error. For BDDA Yu et al. (2020), the model predicts steering and acceleration signals, which are evaluated by mean absolute error. The same trained architecture is used across all benchmarks, with only task-specific input formatting and output heads.

Architecture. We adopt Qwen3-VL Bai et al. (2025) as the base vision-language model. The hidden dimension of the CausalDrive modules is set to 1024. The semantic memory contains 64 slots, and the default prediction horizon is set to 6 for trajectory planning. The reward and cost heads are implemented as lightweight two-layer MLPs with hidden size 512 and dropout 0.1. For constrained GRPO, we sample 4 candidate rollouts per scenario, set the KL coefficient to 0.02, the clipping range to 0.2, and the safety budget to 0.15. The Lagrange multiplier is initialized to 0.05 and updated dynamically during optimization.

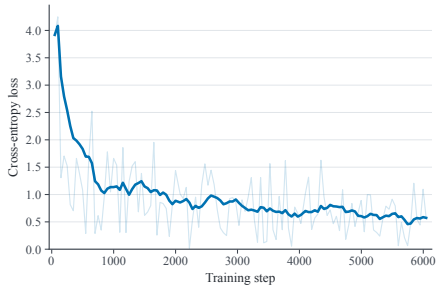
B ADDITIONAL EXPERIMENTAL ANALYSIS

B.1 TRAINING DYNAMICS OF THE TWO-STAGE ALIGNMENT STRATEGY

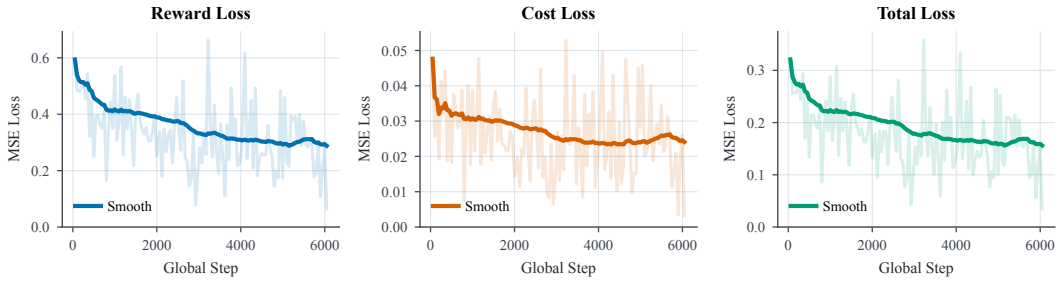
To further analyze the optimization behavior of CausalDrive, we visualize the training dynamics of the two-stage alignment strategy on three benchmarks, including DriveAction Hao et al. (2025), DriveLM Sima et al. (2024), and BDDA Yu et al. (2020). The training process consists of Stage 1, where Supervised Fine-Tuning (SFT) initializes the policy from expert demonstrations, and Stage 2, where learned Reward/Cost Models provide dense feedback signals for constrained Group Relative Policy Optimization (GRPO).

On DriveAction, the training dynamics are shown in Figure S1. During Stage 1, the SFT loss gradually decreases from 4.0 to 0.6, indicating that the policy learns a stable initialization from expert driving behaviors. During Stage 2, the Reward/Cost Models converge steadily and provide stable feedback signals for policy refinement. The adaptive adjustment of the Lagrange multiplier λ during constrained GRPO reflects the progressive incorporation of safety constraints during policy optimization.

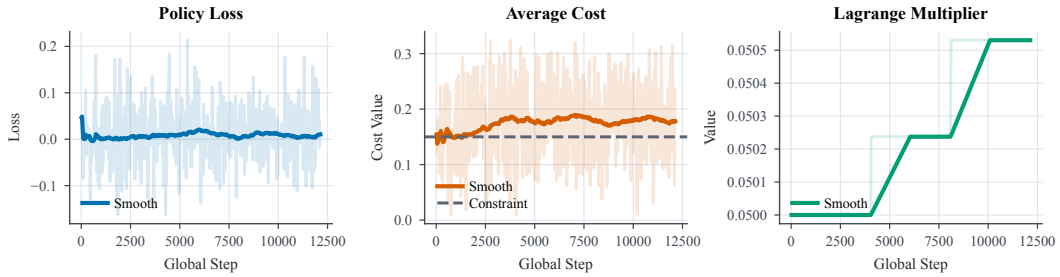
For DriveLM, the training dynamics in Figure S2 demonstrate the optimization process for trajectory planning. During Stage 1, supervised fine-tuning reduces the trajectory L2 error and reaches a stable point at approximately 0.15m. After incorporating learned reward and cost signals, constrained GRPO further refines the policy and reduces the error to 0.05m, corresponding to a 66.7% improvement over



(a) Stage 1: SFT Loss



(b) Stage 2 (Reward): Reward/Cost Modeling



(c) Stage 2 (GRPO): Adaptive λ via GRPO

Figure S1: DriveAction training dynamics across the two-stage alignment process.

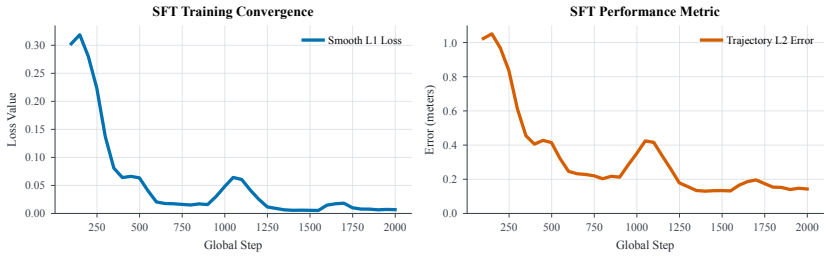
the SFT initialization. This result indicates that the alignment stage provides additional optimization beyond supervised imitation for trajectory planning.

On BDDA, the training dynamics in Figure S3 illustrate the safety-constrained optimization process for continuous control. During the constrained GRPO phase of Stage 2, the safety constraint fulfillment remains consistently below the predefined 0.15 limit. The adaptive adjustment of the penalty strength λ enables the policy to balance control performance and safety requirements during optimization.

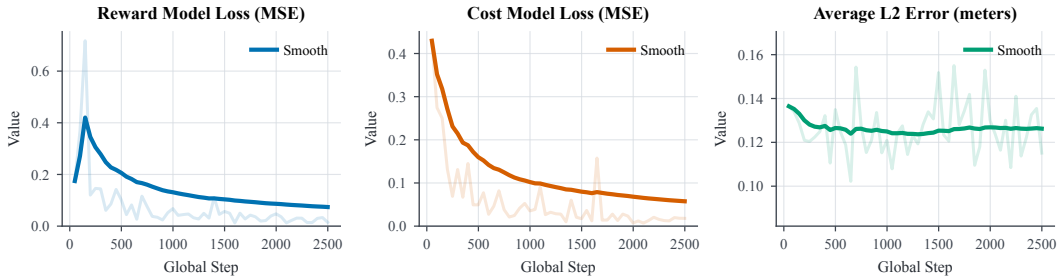
GENERATIVE AI USE DISCLOSURE

In this work, we used generative AI tools to assist with manuscript drafting and editing, improve readability, provide feedback on the presentation of the research methodology and experiments, refine scientific figures. We did not use generative AI tools to generate experimental results or to make scientific conclusions on behalf of the authors.

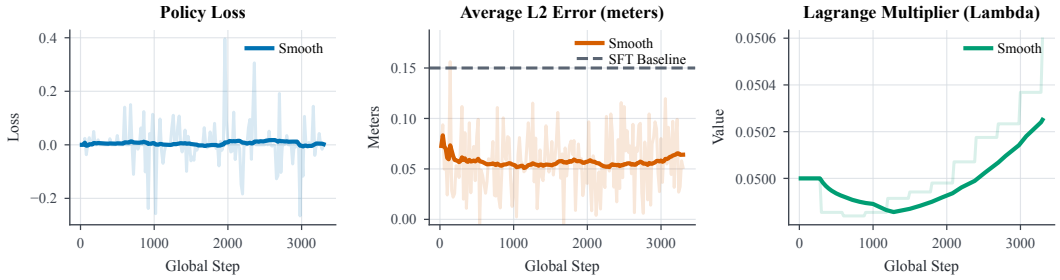
All AI-assisted text, code, figures, and suggestions were carefully reviewed and verified by the authors. Experimental results and conclusions were derived and validated by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.



(a) Stage 1: L2 Error Convergence

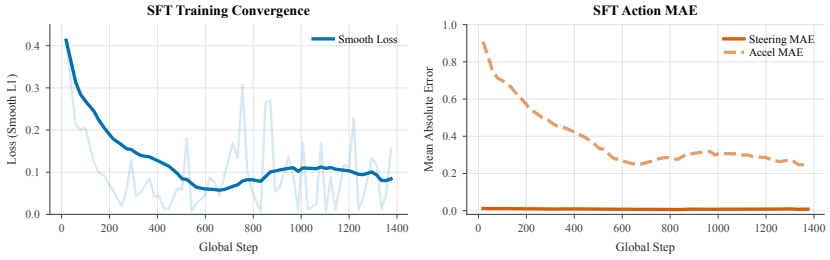


(b) Stage 2 (Reward): Critic MSE Stabilization

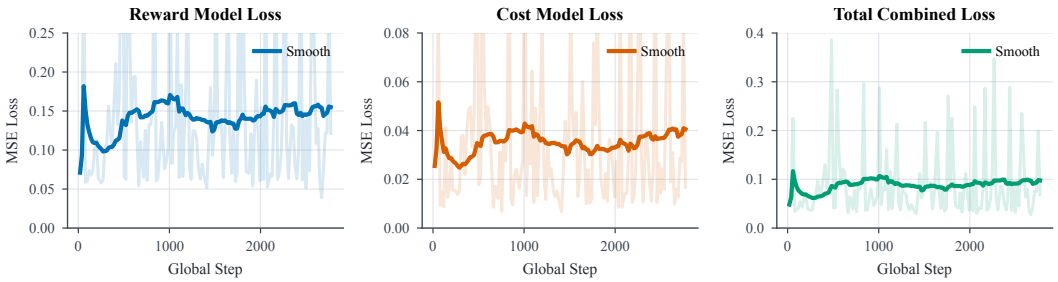


(c) Stage 2 (GRPO): GRPO Trajectory Refinement

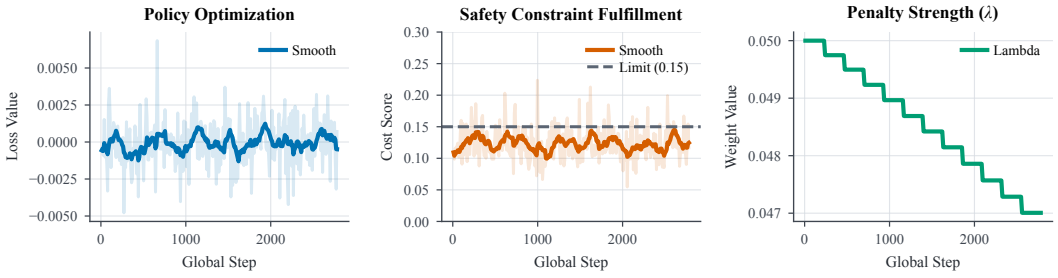
Figure S2: DriveLM training dynamics showing precision improvement beyond the SFT limit.



(a) Stage 1: Control MAE Convergence



(b) Stage 2 (Reward): Critic Combined MSE



(c) Stage 2 (GRPO): Safety Cost Fulfillment

Figure S3: BDDA training dynamics showcasing safe policy alignment under constraints.