2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MAGNIFIED: RL Fine-tuning of Multimodal Large Language Models for Motion Planning

Letian Chen*[1], Yiren Lu*[1], Justin Fu[1], Yichen Xie[1], Runsheng Xu[1], Jyh-Jing Hwang[1], Ben Sapp[1], Drago Anguelov[1]

***Abstract*— Multi-modal Large Language Models (MLLMs) have demonstrated remarkable capabilities in semantic understanding and common sense reasoning, making them promising candidates for solving planning problems in autonomous driving. However, the next-token text prediction objectives traditionally used in pre-training and supervised fine-tuning (SFT) of MLLMs may fall short of fulfilling the planning objectives for autonomous vehicles. The next-token prediction objective merely encourages per-token imitation in text, often irrespective of multi-step consequences and the alignment with crucial planning considerations such as giving space to other road actors. To overcome these limitations, we propose a reinforcement learning fine-tuning (RLFT) approach, MAGNIFIED, that aligns the MLLM-based driving agent with planning objectives by learning from token-level rewards. By mapping a sequence of predicted tokens to corresponding vehicle trajectories and learning from planning rewards, MAGNIFIED optimizes for the true planning objectives rather than focusing solely on token prediction accuracy, enabling the model to refine its understanding of the planning task beyond simple imitation. We validate our approach on the Waymo Open Motion Dataset with a novel setup incorporating rasterized birds-eye views and tokenized trajectories as inputs and planning-oriented outputs. An initial SFT phase establishes a strong baseline in outputting plan trajectories as sequences of X-Y coordinates in text, while subsequent RL fine-tuning substantially enhances planning performance relative to the SFT baseline (demonstrating over a 10.5% reduction in overlap rate and a 38.9% reduction in off-road rate), underscoring the potential of RLFT on MLLMs to achieve vehicle planning that is better aligned with compliant, comfortable, and efficient driving.**

I. Introduction

Multimodal Large Language Models (MLLMs) have demonstrated significant advancements in tasks that require multi-modal (e.g. visual) comprehension and semantic understanding. Leveraging their ability to process both visual and textual inputs, MLLMs can be adapted through Supervised Fine-Tuning (SFT) for a wide range of applications, including visual question answering [1], video captioning [2], medical image analysis [3], robotics [4], [5], and autonomous vehicle (AV) perception [6]. While MLLMs excel in tasks with well-defined answers, their application to AV planning—a domain that demands an understanding of complex scenes, coherent trajectory planning, and decision-making in uncertain environments—remains under-explored [7].

AV planning introduces unique challenges that go beyond traditional visual comprehension tasks. Unlike perception, planning requires generating structured sequences that reflect task-specific constraints and preferences. Despite the potential of MLLMs, the standard SFT paradigm of optimizing next-token prediction may fall short on such tasks, as it only encourages imitation of surface-level token distributions observed in the training data, without regard for whether the resulting trajectories are physically feasible, comfortable, or efficient. This gap between token-level prediction and trajectory-level evaluation mirrors challenges in other domains, such as code generation, where producing executable code cannot be ensured through next-token accuracy alone [8]. In contrast to SFT, Reinforcement Learning (RL) is well-suited towards optimizing task-specific objectives, and has been shown to be effective in guiding models toward performance goals [9], [10]. However, RL training often suffers from high sample complexity, limiting its practical applicability in real-world AV planning, where collecting extensive high-quality data and constructing high-fidelity simulators are both challenging and costly.

To bridge these gaps, we propose a novel approach, MAGNIFIED (reinforceMent leArninG tuNIng For multImodal large languagE moDels), which marries the semantic understanding and common-sense reasoning capabilities of MLLMs with the policy optimization strengths of RL to meet the challenging demands of AV planning. MAGNIFIED leverages the ability of an MLLM base model to interpret a wide distribution of scenarios from extensive pre-training, establishes a strong baseline with planned trajectories as text sequences of X-Y coordinates via a supervised fine-tuning (SFT) phase, and enhances the ability to produce cost-aware plans via an RL Fine-Tuning (RLFT) phase. The SFT phase vastly reduces the exploration space for the RLFT stage, allowing effective targeted improvements in planning objectives with a relatively small amount of compute (1.25% of SFT training steps). MAGNIFIED transforms the MLLM's output text token sequence into a trajectory representation, enabling a planning reward calculation based on trajectory quality on a per-token basis. Another key advantage of MAGNIFIED lies in its compatibility with non-differentiable rewards, which allows it to optimize arbitrary task-specific rewards that improve final end-to-end planning performance. Our contributions are threefold:

1) We introduce MAGNIFIED, an RLFT method designed to enhance MLLMs for AV planning. MAGNIFIED aligns predicted tokens with planning goals by learning from per-token rewards.
2) We apply MAGNIFIED on Waymo Open Motion Dataset (WOMD) with a novel setup for planning:

*Equal Contributions. [1]Waymo LLC.

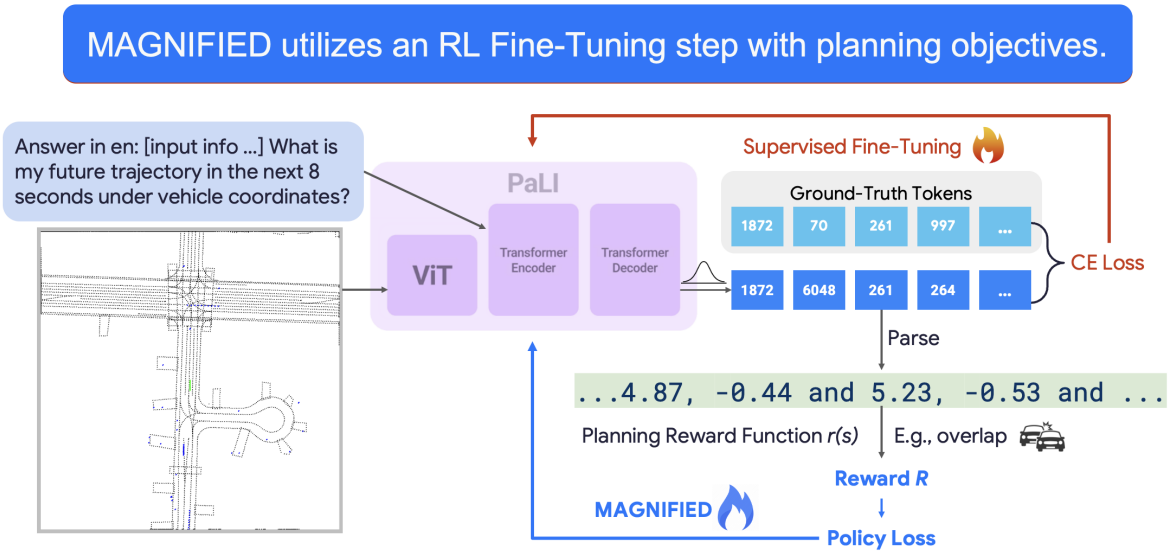

Fig. 1: This diagram illustrates our framework for AV planning. Textual and visual inputs are provided to a MLLM, which generates tokens representing the planned trajectory. Traditional SFT utilizes cross-entropy (CE) losses to optimize next-token prediction. In contrast, MAGNIFIED leverages the semantics of the planned trajectory, calculates planning costs, and applies Reinforcement Learning Fine-Tuning (RLFT) to directly optimize for planning cost.

rasterizing birds-eye view (BEV) roadgraph and road-user information into images, tokenizing trajectories as inputs, and outputting planned trajectories.

3) We demonstrate that MAGNIFIED significantly improves multiple planning objectives compared to SFT: *overlap rate by* 10.5%, and *off-road rate by* 38.9%, while maintaining or *improving* the imitative metrics, particularly in long-horizon planning, even *without* optimizing imitative objectives, transforming MLLMs from imitative predictors into cost-aware planners.

II. Related Work

A. MLLMs for Autonomous Driving

MLLMs have been increasingly explored in autonomous driving to enhance scene understanding and decision-making by integrating visual and textual data. [6] provide a comprehensive survey on MLLM applications in autonomous driving, highlighting their potential in perception, navigation, and control tasks. For example, DriveGPT4 [11], LMDrive [12], Drive Anywhere [13], S4-Driver [14] and OmniDrive [15] utilize LLMs to explain vehicle actions for reasoning and planning. Other recent approaches propose a specialized network structure for detailed contextual understanding [16], or apply chain-of-thought reasoning ([17], [18], [19]). Alpamayo-R1 [20] introduces a structured Chain-of-Causation reasoning framework combined with RL post-training to improve driving reasoning. EMMA [7] presents promising end-to-end motion planning results with SFT. In contrast, our approach, MAGNIFIED, extends the application of MLLMs to AV planning by utilizing RLFT to optimize true planning objectives, addressing the gap in current research work.

B. RL for Autonomous Driving

RL has been widely applied to autonomous driving, focusing on decision-making and control. [21] survey deep RL algorithms for autonomous driving, highlighting challenges such as sample efficiency and the difficulty of collecting datasets that encompass all driving conditions. To address these issues, [22] and [23] improve sample efficiency by incorporating parameterized skills and imitative expert priors, respectively. However, these approaches often rely heavily on large datasets and carefully designed rewards. [24] proposes a neuro-inspired RL framework to enhance safety but notes persistent challenges in generalizing to unseen scenarios. The generalization capabilities of MLLMs, stemming from diverse pre-training data across domains, offer a promising complement to RL. Our method combines the generalization capabilities of MLLMs with the optimization strengths of RL, enabling a more cost-aware planning system.

C. RL Fine-Tuning for LLMs

RLFT has been explored across various domains to enhance LLMs by aligning their outputs with specific objectives. For instance, [25] use RL from human feedback to fine-tune language models, improving their alignment with human preferences. [26] explore RLFT to mitigate harmful outputs, demonstrating the effectiveness of RL in refining model behaviors. In the context of vision-language tasks, [27] apply RLFT to improve image captioning models by improving generated captions on human evaluations. [28] applies RLFT on vision-language models in order to solve a variety of simulated games. Rather than fine-tuning a model directly, [29] utilize a MLLM to generate a reward model that can be optimized using RL. Despite these advancements, most applications focus on tasks with well-defined outputs, such as image captions or executable codes. In contrast,

AV planning presents uncertain, sequential decision-making challenges. MAGNIFIED extends RLFT to MLLMs in this domain by introducing token-level rewards, enabling precise credit assignment to optimize planning objectives.

III. Preliminaries

A. Markov Decision Process

A Markov Decision Process (MDP) is defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. $\mathcal{S}$ is the set of states, and $\mathcal{A}$ is the set of actions the agent can take. The transition probability $P(s'|s,a)$ represents the likelihood of reaching a state s' from state s when action a is taken. The reward function $r(s)$ provides a scalar feedback signal at each state, and $\gamma \in [0,1]$ is the temporal discount factor. The goal in RL is to learn a policy π that maximizes the expected return, $V(s_t) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k})]$, starting from state s_t. In AV planning, this translates to training a policy that plans future waypoints as actions based on the current state of the environment.

B. REINFORCE with KL Penalty

The REINFORCE algorithm is a classic policy-gradient method that learns a policy by maximizing the expected return via stochastic gradient ascent [30], as shown in Equation 1.

$$\max_{\pi_\theta} \mathbb{E}_{(s_t,a_t)\sim\pi_\theta} \left[\sum_{t=0}^{T} \log \pi_\theta(a_t|s_t) \tilde{Q}_{\pi_\theta}(s_t,a_t) \right] \quad (1)$$

The state-action value function, Q, is typically estimated using the empirical returns, $\tilde{Q}_\pi(s_t,a_t) = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$, where $r_{t'}$ is the empirical reward obtained at timestep t' with a rollout starting with (s_t,a_t) and following policy π afterwards. π_θ is a policy parameterized by θ. A common improvement of the vanilla REINFORCE algorithm is to introduce a learned value baseline, V_ϕ, to lower the variance of return estimation while maintaining unbiased. We use L2 loss to learn the value baseline: $l_{V_\phi} = ||V_\phi(s_t) - \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}||^2$. For fine-tuning large language models, it is also common to include a Kullback–Leibler (KL)-divergence regularization between the reference model and the fine-tuned model [31]. This augments the final objective to:

$$\max_{\pi_\theta} \mathbb{E}_{(s_t,a_t)\sim\pi_\theta} \Big[(1-\alpha) \sum_{t=0}^{T} \log \pi_\theta(a_t|s_t) (\tilde{Q}_{\pi_\theta}(s_t,a_t) - V_\phi(s_t)) - \alpha D_{\text{KL}}(\pi_\theta(\cdot|s_t)||\pi_{\text{ref}}(\cdot|s_t)) \Big] \quad (2)$$

where α controls the relative weight for the KL regularization, and π_{ref} is the base reference policy.

Normalization of advantages. The advantage estimate, $\tilde{A}_\pi$, is defined as $\tilde{A}_\pi(s,a) = \tilde{Q}_\pi(s,a) - V_\phi(s)$. We follow standard practice to normalize the advantage estimates across a batch of transitions: $A_\pi = (\tilde{A}_\pi - \text{mean}(\tilde{A}_\pi))/\text{std}(\tilde{A}_\pi)$.

C. PaLI-3: Vision-Language Models

In this work, we utilize PaLI-3 [32], [33], as the MLLM. PaLI-3 has 5B parameters, distributed between a 2B contrastively pre-trained Vision Transformer (ViT) and a 3B fused encoder-decoder [34]. The ViT takes images as input and produces vision embeddings. The fused encoder-decoder takes the vision embeddings and language tokens to predict language tokens in an auto-regressive fashion. PaLI-3 has shown competitive performance across various multimodal benchmarks and computational efficiency suitable for fine-tuning [33].

IV. Method: MAGNIFIED

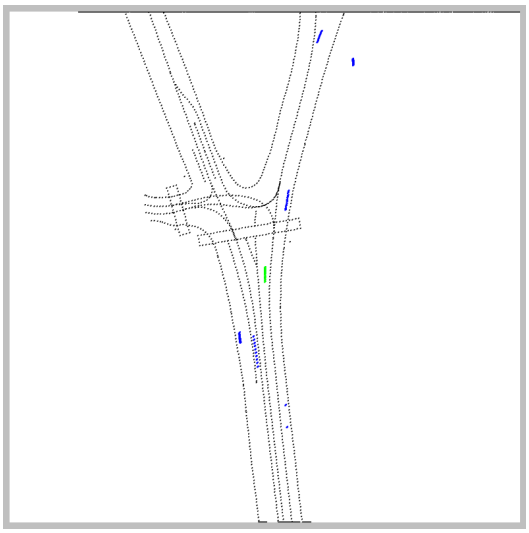

Fig. 2: Example of rasterized image. The roadgraph is shown in black for better visibility.

We present MAGNIFIED, a Reinforcement Learning Fine-Tuning (RLFT) framework that enhances MLLMs for AV planning. This section outlines the problem setup, input representation, reward formulation, and the reinforcement learning used in MAGNIFIED.

A. Problem Setup

We formulate the AV planning task as a trajectory generation problem where the planner predicts the future trajectory of the ego-vehicle given the scene context, including past trajectories of other agents and the roadgraph. Trajectories are represented as a sequence of (x,y) waypoints. In this work, we use the WOMD [35], which provides rich data for planning scenarios. Each scenario involves predicting an 8-second trajectory for the ego-vehicle based on 1 second of historical information. The dataset includes information on the location, speed, and acceleration of the ego-vehicle and other agents, along with roadgraph information. The model is tasked to output the predicted trajectory as a sequence of (x,y) waypoints. The data frequency is 10Hz, meaning that each input sequence consists of 10 waypoints for the past second, and the model must generate 80 waypoints.

B. Visual and Textual Input Representation

A key challenge in MLLM planning is designing effective input modalities. We propose a novel visual and textual representation tailored for trajectory planning. To encode the driving scene, we rasterize the roadgraph and past trajectories

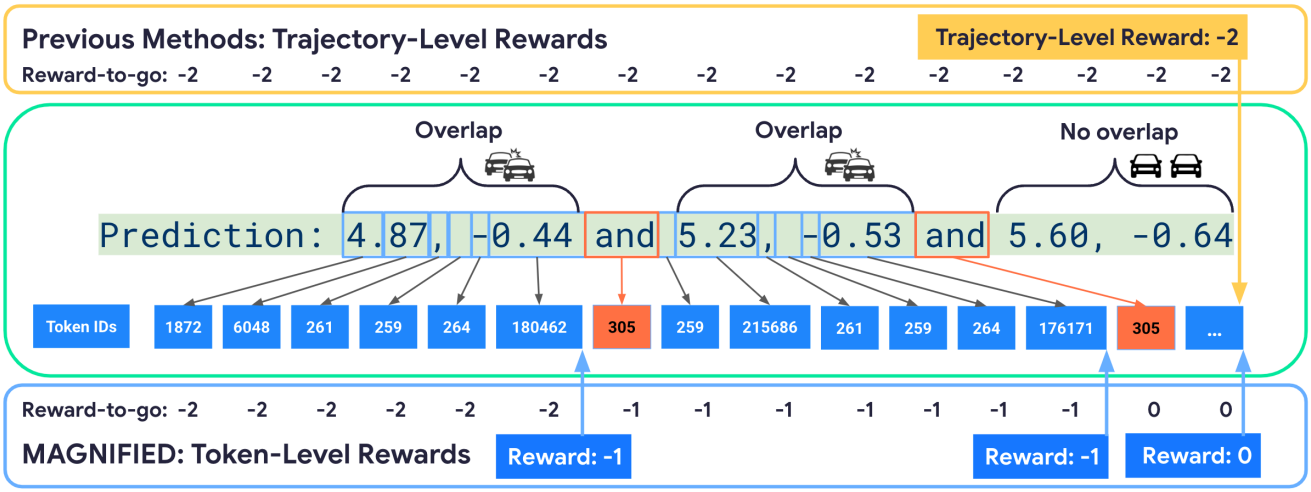

Fig. 3: This figure illustrates the comparison between token-level reward and sequence-level reward. In this example, the model predicts three waypoints: (4.87, -0.44), (5.23, -0.53), and (5.60, -0.64). The corresponding token IDs are displayed in the blue boxes. Previous methods calculate the total overlap for the entire trajectory and assign a single reward at the end, resulting in a constant reward-to-go. In contrast, MAGNIFIED assigns rewards at the token level by associating each reward with the token preceding an "and" token (e.g., token 305). This approach captures the impacts of key tokens on the objective, facilitating more precise credit assignment.

into a 3-channel (RGB) image with shape $[1064, 1064, 3]$. An example is shown in Fig 2. Each channel conveys distinct scene elements: **Red**: represents the roadgraph, including lane boundaries and other road features. **Green**: encodes the past 1-second trajectory of the ego-vehicle. **Blue**: displays the past 1-second trajectories of other vehicles in the scene.

All coordinates are transformed to center the ego-vehicle at $(0, 0)$ and its orientation facing upward. This concise representation captures the traffic flow, relative positions, and recent motion history in a form that MLLMs can intuitively interpret. We find that PaLI-3's ViT encoder effectively processes this rasterized BEV input even without fine-tuning.

In addition to the visual input, textual inputs include ego-vehicle's position, velocity, and acceleration over the past second. We also include route [36], which consists of the logged future trajectory projected to the roadgraph points. We sample 20 points from the route with a 5-point interval. The projection and sampling minimizes the information leaked about the future path. The off-the-shelf PaLI tokenizer was used to encode the text and allow MAGNIFIED to leverage the natural language knowledge prior from PaLI.

One example of the textual input is shown below:

```
Answer in en: Assume I am at the coordinate
0,0.The high-level behavior attention is: go
follow route:-1.71, -0.16 and 3.27, -0.13...
The past trajectory under vehicle coordinate
is: -2.21, 0.00 and -1.93, 0.00... The past
ego velocity under vehicle coordinate is:
2.75, -0.00 and 2.61, -0.00... The past ego
acceleration under vehicle coordinate is:
-1.21, 0.02 and -1.33, -0.01... Other agent
current locations under vehicle coordinate
is: 27.81, 3.80 and -11.59, -0.61... What is
my future trajectory in next 8 seconds under
vehicle coordinate?
```

One example of the output trajectory is:

```
0.14, -0.00 and 0.27, -0.00 and 0.39, -0.00
and ...
```

C. Token-level Rewards

To align the MLLM-based planner with planning objectives, we propose token-level rewards, facilitating more precise credit assignment. In this work, we focus on two core planning objectives: 1) avoiding overlaps with other actors, and 2) avoiding driving off-road.

MAGNIFIED evaluates the cost of each predicted waypoint by checking for overlaps with other agents and whether the agent is off-road. The reward function is defined as a combination of the overlap counts and the off-road indicator by a coefficient, w_o: in Equation 3, c_t^{overlap} denotes the number of overlapped agents at time t, and $\mathbb{I}_t^{\text{off-road}}$ denotes whether the vehicle is off-road.

$$r_t = -(1 - w_o)c_t^{\text{overlap}} - w_o\mathbb{I}_t^{\text{off-road}} \tag{3}$$

Unlike typical RLFT in LLMs which often rely on sequence-level rewards, MAGNIFIED leverages the structure of trajectory outputs to provide dense supervision. The predicted trajectory is represented as a sequence of (x, y) waypoints separated by the token "and", and rewards are assigned to the token preceding each "and" (Figure 3). This structure facilitates finer-grained credit assignment by linking tokens to planning outcomes, such as overlap or off-road violations.

D. Policy Gradient with KL Penalty

At the core of MAGNIFIED is a RLFT framework that leverages token-level rewards to align the MLLM with cost-aware planning objectives. We choose to utilize REINFORCE augmented with a KL-divergence penalty to optimize the planning objectives while maintaining output's trajectory structure and preventing catastrophic forgetting.

Model	ADE@8s↓	Overlap Rate↓	Overlap Count↓	Off-Road Rate↓	Route Progress
Baselines					
Wayformer (re-plan) [36]†	N/A	10.68%	N/A	7.89%	123.58%
SFT	1.785	10.10%	1.71	5.60%	94.9%
SFT w/ blackout images	3.179	18.9%	3.29	7.80%	105.7%
SFT w/o route	1.787	10.7%	1.78	6.64%	96.4%
MAGNIFIED (Ours)					
$w_o = 0.0$ (Overlap only)	1.712	8.76%	1.48	5.88%	100.5%
$w_o = 0.25$	**1.704**	**8.60%**	**1.41**	3.68%	103.1%
$w_o = 0.5$	1.726	9.04%	1.51	**3.42%**	100.9%
$w_o = 0.75$	1.722	9.16%	1.54	3.56%	98.6%
$w_o = 1.0$ (Off-road only)	1.770	10.06%	1.70	3.57%	95.4%

TABLE I: Comparisons between baselines, SFT, and MAGNIFIED (ours) on planning metrics. **Overlap Rate** and **Off-Road Rate** are binary for each run segment and averaged over dataset. **Overlap Count** represents the averaged number of overlap instances in each run segment. † Wayformer (re-plan) uses the same route as MAGNIFIED and re-plans every 5 steps [36].

MAGNIFIED first parses the model's output tokens as a sequence of (x, y) waypoints separated by the token "and". The reward-to-go, $\tilde{Q}$, is then calculated and used to compute the policy gradient (Equation 2). By associating rewards with the corresponding tokens, MAGNIFIED captures the cumulative impact of tokens on the trajectory quality. While our implementation of MAGNIFIED focuses on overlap and off-road avoidance, the framework is adaptable to other planning objectives.

V. Results

In this section, we present the experimental results on the Waymo Open Motion Dataset (WOMD) [35], which includes 483,433 training samples and 43,783 evaluation samples. All experiment hyper-parameters are summarized in Table IV. Our experiments are designed to answer the following questions:

Q1: Can the SFT stage take in our proposed visual and textual inputs (Sec. IV-B) and enable PaLI-3 to predict structured trajectories that achieve decent imitative performance?

Q2: Can MAGNIFIED improve *targeted planning objectives* compared to the SFT baseline?

Q3: Can MAGNIFIED simultaneously improve *multiple planning objectives*?

Q4: Is token-level reward more effective than traditional sequence-level rewards?

Q5: Do KL regularization and other hyper-parameters impact RLFT performance?

A. Benchmark Results

We evaluate MAGNIFIED and SFT on both imitation and planning performance metrics. SFT models are trained for 100,000 steps with a batch size of 256. RLFT models are trained for 10,000 steps with a batch size of 32. Results are reported on the evaluation set.

Trajectory Prediction Metrics. Table II reports Average Displacement Error (ADE) and Final Displacement Error (FDE) at 3, 5, and 8 seconds. ADE@Ns represents the average L2 distance between the prediction and ground-truth ego-vehicle positions over the horizon of N seconds. FDE@Ns represents the final L2 distance after N seconds has elapsed. All baselines, including MotionLM [37], Wayformer [38] and EMMA [7], sample multiple trajectories (24-192 samples), which are subsequently aggregated clustering into the final trajectory, whereas our approaches only generates one trajectory. Nevertheless, SFT achieves comparable imitative performance, suggesting it successfully enables PaLI-3 to process rasterized image inputs and text instructions, and generate trajectories in the structured format, establishing a strong baseline for evaluation. In particular, the significant degradation of SFT w/ blackout images (e.g., 78% increases on ADE@8s) highlights the importance of the visual input.

Surprisingly, although MAGNIFIED does not explicitly optimize ADE or FDE, it achieves better long-horizon imitative accuracy: MAGNIFIED ($w_o = 0.25$) achieves a 4.5% ADE@8s improvement and a 8.0% FDE@8s improvement over SFT, as shown in Figure 4. Performance at 3s and 5s remains comparable. We hypothesize that optimizing for key planning objectives captures fundamental driving intents, resulting in improved long-horizon imitative performance.

Planning Metrics. Table I presents results for planning metrics: Overlap Rate, Overlap Count, Off-Road Rate, and Route Progress, following [36]. These metrics evaluate the quality of the generated trajectory, beyond mere imitation. When optimizing only for overlap avoidance, MAGNIFIED reduces Overlap Rate from 10.10% (SFT) to 8.76% (a 13.3% improvement) and Overlap Count from 1.71 to 1.48 (a 13.5% improvement), confirming its ability to learn overlap-avoidance behavior. When optimizing only for off-road avoidance, MAGNIFIED reduces off-road rate from 5.60% to 3.57%, a 36.3% reduction. These results confirm that MAGNIFIED can successfully optimize for specific planning objectives, such as overlap and offroad, while keeping the behavior neutral as shown in ADE@8s. Notably, these improvements are achieved using only 1.25% training steps of SFT, demonstrating the sample efficiency of MAGNIFIED.

Route Information. To evaluate the contribution of route input, we compare "SFT" with "SFT w/o route" in Table I and Table II. The "SFT w/o route" variant instead provides a

Model	ADE@3s ↓	ADE@5s ↓	ADE@8s ↓	FDE@3s ↓	FDE@5s ↓	FDE@8s ↓
Baselines*						
MotionLM [37]	0.251	0.694	1.766	N/A	N/A	N/A
Wayformer [38]	0.250	0.640	1.517	N/A	N/A	N/A
EMMA [7]	0.248	0.681	1.718	N/A	N/A	N/A
MAGNIFIED (Ours)						
Overlap only $w_o = 0.0$	0.244	0.695	1.712	0.726	2.042	4.812
$w_o = 0.25$	0.251	0.704	1.704	0.743	2.047	4.714
$w_o = 0.5$	0.252	0.709	1.726	0.746	2.065	4.813
$w_o = 0.75$	0.250	0.705	1.722	0.740	2.056	4.823
Off-road only $w_o = 1.0$	0.252	0.715	1.770	0.749	2.101	5.016
Supervised Tuning						
SFT	0.245	0.709	1.785	0.735	2.106	5.124
SFT w/ blackout images	0.469	1.260	3.179	1.335	3.661	9.334
SFT w/o route	0.248	0.711	1.787	0.739	2.103	5.126

TABLE II: Comparison between baselines, SFT, and MAGNIFIED (ours) on imitative metrics in Waymo Open Motion Dataset. * Baseline ADE results reported in [7].

coarse high-level navigation command (go straight/left/right). Although this version still produces reasonable trajectories, we observe slight improvements across all metrics when route information is included. The close performance between the two conditions suggests that the MLLM-planning paradigm is robust to degraded input information.

B. Rewards Trade-off Analysis

To answer **Q3**, we evaluate whether MAGNIFIED can simultaneously optimize multiple planning objectives—specifically, reducing both overlap and off-road events. We vary the reward weight w_o in Equation 3, which controls the emphasis on minimizing off-road instead of overlap. Figure 5 illustrates how varying w_o affects the Overlap Rate and Off-road Rate. As w_o increases, the Overlap Rate generally increases while the Off-road Rate decreases. On planning metrics (Table I), MAGNIFIED successfully improves both overlap and off-road metrics when using mixed-objective rewards. For example, with $w_o = 0.5$, MAGNIFIED reduces Overlap Rate from 10.10% (SFT) to 9.04% (a 10.5% improvement), and Off-road Rate from 5.60% to 3.42% (a 38.9% reduction). Notably, the mixed setting $w_o = 0.25$ further reduces the Overlap Rate to 8.60% (a 14.9% improvement) and the Off-road Rate to 3.68% (a 34.3% improvement). In contrast, when optimizing only for one objective, MAGNIFIED improves that metric while slightly compromising the other. For example, overlap-only reward reduces Overlap Rate but results in a slightly higher Off-road Rate of 5.88%. These findings confirm that MAGNIFIED not only is able to optimize individual planning objectives, but also is able to improve multiple objectives simultaneously, all while preserving imitative performance.

C. Ablation Studies

We conduct ablations to assess the importance of design choices and answer **Q4** and **Q5**.

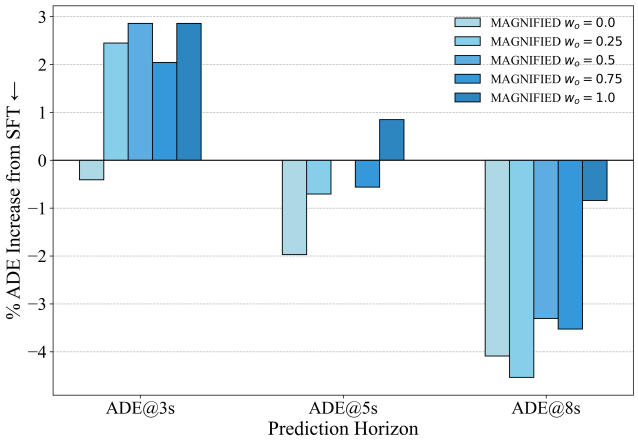

Fig. 4: Percentage change in ADE at 3, 5, and 8 seconds for MAGNIFIED relative to SFT. Negative values indicate improvements over SFT. MAGNIFIED achieves better long-horizon imitation *without* explicitly optimizing ADE.

Model	ADE@8s	Overlap Rate	Overlap Count
$\alpha = 0.0$ (w/o KL)	1.872	9.76%	1.57
$\alpha = 0.1$	1.712	**8.76%**	**1.48**
$\alpha = 0.3$	**1.704**	8.79%	1.49
$\alpha = 0.5$	1.749	9.76%	1.63
$\alpha = 0.1$, w/o Token-Rew	1.760	9.92%	1.69

TABLE III: MAGNIFIED ablation results. All runs use Overlap only reward ($w_o = 0.0$).

Token-Level Reward. We compare token-level MAGNIFIED against a variant that uses sequence-level rewards—i.e., summing rewards across the trajectory as a single scalar reward. This ablation, reported in Table III (row "w/o Token-Rew"), isolates the contribution of our token-level reward assignment. The results suggest that token-level rewards enable more effective learning and improve planning performance compared with sequence-level rewards.

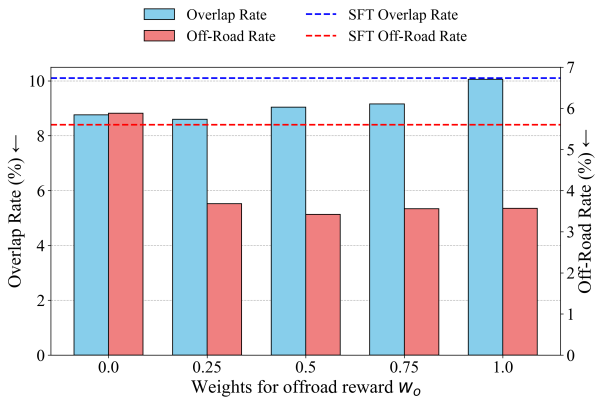

Fig. 5: Effect of the off-road weight w_o on planning metrics. As w_o increases, overlap and off-road trade off, confirming MAGNIFIED's ability to balance multiple objectives.

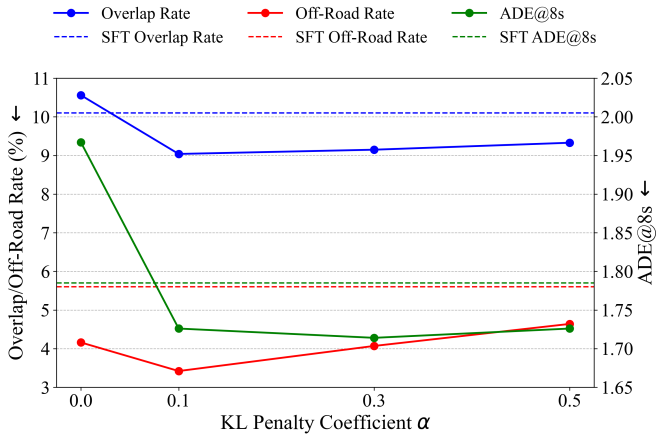

Fig. 6: This figure shows the impact of KL regularization weight α on planning metrics ($w_o = 0.5$). Moderate regularization ($\alpha = 0.1$) leads to the best performance in both Overlap Rate and Off-road Rate, while too little or too much regularization degrades performance.

RL Hyper-parameters. We conduct experiments with different values of the KL regularization weight $\alpha \in \{0.1, 0.3, 0.5\}$, which controls the strength of penalty for divergence from the reference policy. We show results in Table III and Figure 6. While all variants lead to improvements over the SFT baseline, we find that $\alpha = 0.1$ achieves the best performance. This supports the intuition that KL regularization serves as a stabilizing auxiliary term, while overly strong penalties may limit performance.

KL Penalty. To further assess the role of the KL penalty, we ablate the KL loss entirely ("w/o KL" in Table III and $\alpha = 0.0$ in Figure 6). Removing the KL penalty results in degraded performance across both planning and imitation metrics, highlighting its importance in stabilizing learning and preserving imitative behaviors.

D. Qualitative Analysis

We present examples illustrating how MAGNIFIED resolves overlap instances from SFT in the supplementary video. In the Overlap Case 1, the SFT-controlled ego-vehicle moves too quickly when another vehicle merges into its lane. At $t = 3s$, the ego-vehicle overlaps with the orange vehicle. MAGNIFIED, however, adjusts its speed and yields to the merging vehicle, demonstrating better planning and awareness of the merging vehicle's trajectory. More qualitative examples in the video demonstrate how MAGNIFIED effectively addresses overlap and offroad risks with RLFT on planning objectives.

Hyper-parameter	Value / Scheme
SFT learning rate	linear warm-up from 0 to 3e-3 in 1,000 steps, then delay with slope 1e-4 / $\sqrt{\text{step}/1000}$
SFT batch size	256
SFT training steps	100,000
RLFT learning rate	linear warm-up from 0 to 1e-4 in 2,000 steps, then constant 1e-4
RLFT batch size	32
RLFT training steps	10,000
MDP temporal discount	1.0

TABLE IV: Hyper-parameters and their values.

VI. Conclusion

In this work, we present MAGNIFIED, an RLFT framework that transforms MLLMs into cost-aware autonomous driving planners. Our approach leverages MLLMs' semantic understanding and common-sense reasoning capabilities, and an RLFT phase integrating novel token-level rewards to directly optimize for planning objectives. Experiments on WOMD demonstrate significant overlap reduction ($> 10\%$) and off-road reduction ($> 38\%$) compared to baseline SFT, while maintaining or improving the imitative behaviors.

Future work could incorporate additional reward signals, enabling MAGNIFIED to address a broader range of planning objectives and real-world constraints. Another line of future work is to test MAGNIFIED beyond AV planning, such as robotics, where multi-turn RL with per-token rewards is suitable to enable long-horizon planning capabilities.

VII. Limitations

While MAGNIFIED demonstrates strong improvements in planning metrics, several limitations remain. First, our rewards focus only on overlap and off-road penalties, whereas real-world planning involves broader objectives such as comfort and road rules compliance. Second, MAGNIFIED relies on a specific token structure for reward assignment, which may limit generalization to other output formats. Third, our experiments are conducted exclusively on WOMD; evaluating transferability to other domains is left for future work. Finally, all evaluations of MAGNIFIED are performed open-loop – closed-loop evaluation is important to assess compounding errors and interactive behaviors. We expect that reductions in overlap and off-road deviations at the trajectory level would translate to improved performance in closed-loop simulation. However, closed-loop effects such as

distribution shift and multi-agent interactions may introduce additional challenges.

REFERENCES

[1] U. Naseem, M. Khushi, and J. Kim, "Vision-language transformer for interpretable pathology visual question answering," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1681–1690, 2022.

[2] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 714–10 726.

[3] Y. Bazi, M. M. A. Rahhal, L. Bashmal, and M. Zuair, "Vision–language model for visual question answering in medical imagery," *Bioengineering*, vol. 10, no. 3, p. 380, 2023.

[4] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 462–12 469.

[5] C. Xu, H. T. Nguyen, C. Amato, and L. L. Wong, "Vision and language navigation in the real world via online visual language mapping," *arXiv preprint arXiv:2310.10822*, 2023.

[6] X. Zhou, M. Liu, E. Yurtsever, B. L. Zagar, W. Zimmer, H. Cao, and A. C. Knoll, "Vision language models in autonomous driving: A survey and outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.

[7] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, *et al.*, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024.

[8] S. Dou, Y. Liu, H. Jia, L. Xiong, E. Zhou, W. Shen, J. Shan, C. Huang, X. Wang, X. Fan, *et al.*, "Stepcoder: Improve code generation with reinforcement learning from compiler feedback," *arXiv preprint arXiv:2402.01391*, 2024.

[9] Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson, *et al.*, "Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7553–7560.

[10] Z. Peng, W. Luo, Y. Lu, T. Shen, C. Gulino, A. Seff, and J. Fu, "Improving agent behaviors with rl fine-tuning for autonomous driving," in *European Conference on Computer Vision*. Springer, 2025, pp. 165–181.

[11] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *RA-L*, 2024.

[12] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," in *CVPR*, 2024.

[13] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, and D. Rus, "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," in *ICRA*, 2024.

[14] Y. Xie, R. Xu, T. He, J.-J. Hwang, K. Luo, J. Ji, H. Lin, L. Chen, Y. Lu, Z. Leng, *et al.*, "S4-driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1622–1632.

[15] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024.

[16] C. Pan, B. Yaman, T. Nesti, A. Mallik, A. G. Allievi, S. Velipasalar, and L. Ren, "Vlp: Vision language planning for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 760–14 769.

[17] X. Tian, J. Gu, B. Li, Y. Liu, C. Hu, Y. Wang, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," in *CoRL*, 2024.

[18] T. Wang, E. Xie, R. Chu, Z. Li, and P. Luo, "Drivecot: Integrating chain-of-thought reasoning with end-to-end driving," *arXiv preprint arXiv:2403.16996*, 2024.

[19] A. Bhattacharyya, S. Panchal, M. Lee, R. Pourreza, P. Madan, and R. Memisevic, "Look, remember and reason: Grounded reasoning in videos with language models," in *ICRA*, 2023.

[20] L. Ullrich, M. Buchholz, K. Dietmayer, and K. Graichen, "Toward fully autonomous driving: Ai, challenges, opportunities, and needs," *arXiv preprint arXiv:2601.22927*, 2026.

[21] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2021.

[22] L. Wang, J. Liu, H. Shao, W. Wang, R. Chen, Y. Liu, and S. L. Waslander, "Efficient reinforcement learning for autonomous driving with parameterized skills and priors," *arXiv preprint arXiv:2305.04412*, 2023.

[23] Z. Huang, J. Wu, and C. Lv, "Efficient deep reinforcement learning with imitative expert priors for autonomous driving," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7391–7403, 2022.

[24] X. He, J. Wu, Z. Huang, Z. Hu, J. Wang, A. Sangiovanni-Vincentelli, and C. Lv, "Fear-neuro-inspired reinforcement learning for safe autonomous driving," *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[25] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[26] J. Wang, H. Wang, S. Sun, and W. Li, "Aligning language models with human preferences via a bayesian approach," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[28] S. Zhai, H. Bai, Z. Lin, J. Pan, P. Tong, Y. Zhou, A. Suhr, S. Xie, Y. LeCun, Y. Ma, *et al.*, "Fine-tuning large vision-language models as decision-making agents via reinforcement learning," *Advances in neural information processing systems*, vol. 37, pp. 110 935–110 971, 2024.

[29] Y. Wang, Z. Sun, J. Zhang, Z. Xian, E. Biyik, D. Held, and Z. Erickson, "Rl-vlm-f: Reinforcement learning from vision language foundation model feedback," *arXiv preprint arXiv:2402.03681*, 2024.

[30] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[31] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker, "Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms," *arXiv preprint arXiv:2402.14740*, 2024.

[32] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, *et al.*, "Pali-x: On scaling up a multilingual vision and language model," *arXiv preprint arXiv:2305.18565*, 2023.

[33] X. Chen, X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, B. Mustafa, S. Goodman, I. Alabdulmohsin, P. Padlewski, *et al.*, "Pali-3 vision language models: Smaller, faster, stronger," *arXiv preprint arXiv:2310.09199*, 2023.

[34] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[35] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[36] C. Gulino, J. Fu, W. Luo, G. Tucker, E. Bronstein, Y. Lu, J. Harb, X. Pan, Y. Wang, X. Chen, *et al.*, "Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

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

[38] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2980–2987.