MODEL-AGNOSTIC CONTRASTIVE ADAPTATION FOR SLOW-FAST END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

The dual-system architecture combining fast-reactive end-to-end (E2E) models with slow-reasoning Vision Language Models (VLMs) has emerged as a promising solution for autonomous driving in complex scenarios. However, the rapid iteration of E2E models demands a flexibly decoupled, model-agnostic framework. While injecting latent features from VLMs into E2E models preserves dense semantic guidance and inherently offers a higher theoretical upper bound for performance, naive implementations often fail, forcing researchers to adopt intricate, deeply coupled fusion structures. In this paper, we reveal that this plugin failure is not a structural deficiency but an optimization trap. Due to cross-model shortcut learning and gradient vanishing, pre-trained E2E models instinctively treat injected semantic priors as perturbation noise. To address this, we propose MACA, a novel model-agnostic cognitive plugin. MACA bypasses high-dimensional feature injection by extracting concise semantic intentions via a two-stage Sequential Reasoning-to-Decision Training (SRDT) strategy. To break optimization stagnation, we introduce Relative Advantage Preference Optimization (RAPO), which evaluates the VLM-augmented model against a frozen E2E baseline to generate dynamic comparative signals, compelling the network to deeply internalize VLM cognition. Extensive evaluations demonstrate MACA's generality across diverse state-of-the-art E2E architectures (*e.g.*, DiffusionDrive, SparseDrive, MomAD, GoalFlow). For instance, integrating MACA into DiffusionDrive reduces the average L_2 error from $0.57\,\mathrm{m}$ to $0.51\,\mathrm{m}$ and the collision rate from 0.08% to 0.05% on nuScenes, while improving the PDMS of DiffusionDrive and GoalFlow from 87.5 and 89.2 to 88.2 and 90.0, respectively, on the NAVSIM benchmark. Our approach ensures driving stability in regular scenarios while significantly boosting decision-making success in corner cases, comprehensively outperforming existing baselines. Code and models will be released.

1 INTRODUCTION

The paradigm of end-to-end (E2E) autonomous driving has successfully unified perception and planning for efficient and reactive vehicle control. However, navigating complex unscripted scenarios demands advanced reasoning capabilities. While pure Vision Language Action (VLA) models attempt to address this requirement (Yang et al., 2025b), they typically suffer from high inference latency and prohibitive training costs that hinder real-time deployment. In the recent DrivingBench real-vehicle experiment, GPT-6 Astra completed the full course while performing deliberative reasoning over visual observations and issuing motion commands at discrete intervals. During model inference, the vehicle-side control stack continued executing the issued commands (Ramabadran et al., 2026). This setup illustrates a clear division between slow reasoning and fast execution. It also suggests that physical intelligence need not rely on a single model to handle both complex cognition and real-time control. Instead, combining the high-level understanding of a VLM with a fast-reactive E2E model may offer a more practical and attainable path. Since E2E models iterate significantly faster than VLMs in both academia and industry, such a dual-system architecture requires a model-agnostic and decoupled framework. This design allows researchers and engineers to upgrade to the latest E2E architectures without being constrained by tightly integrated systems.

The key to realizing this framework lies in establishing effective communication between the VLM and the E2E driving model. Early strategies primarily relied on generating explicit future way-

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.

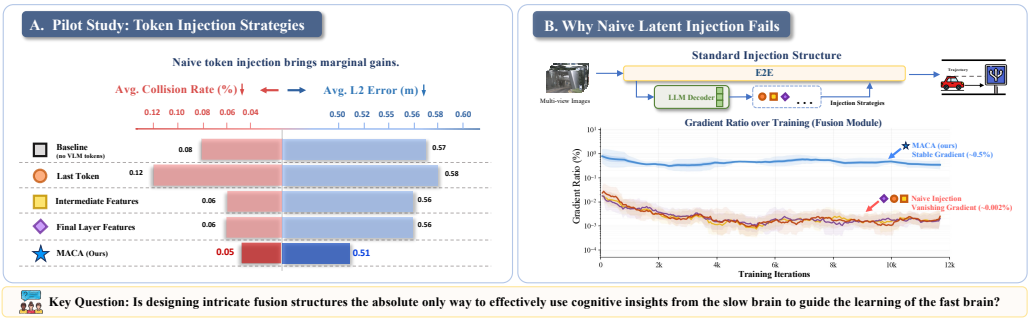


Figure 1: **The optimization trap in VLM-E2E fusion. (A)** Naive VLM latent injection (last-token, intermediate, or final-layer features) yields only marginal gains or even degradation. The main factor is the training objective: without preference-driven optimization, changing which VLM features are injected brings little benefit. **(B)** We define the gradient ratio as $\|g_{\text{fusion}}\|_2 / \|g_{\text{all}}\|_2$, where g_{fusion} aggregates gradients of the injection-module parameters and g_{all} aggregates all trainable parameters (excluding frozen VLM weights). Under naive injection this ratio is $\sim$0.002%; MACA maintains $\sim$0.5%, mitigating gradient vanishing.

points (Tian et al., 2024; Chen et al., 2025) or compressing decisions into discrete (Jiang et al., 2024; Gariboldi et al., 2025; Qian et al., 2024). These two paradigms create a clean boundary between systems, making it incredibly straightforward to decouple the fast reactive brain from the slow-reasoning brain for independent upgrades. Yet, a fundamental contradiction remains. While waypoint generation seems straightforward for planner initialization, language models are inherently insensitive to continuous spatial coordinates. Forcing such geometric outputs neglects their semantic reasoning strengths. Similarly, reducing decisions to one-hot discrete actions creates a severe information bottleneck that eradicates rich contextual semantics.

To overcome these limitations, recent research explores injecting the E2E model with either intermediate features from the VLM or newly encoded embeddings derived from its linguistic outputs that contain rich context and explicit decisions. This representation-level injection preserves dense semantic guidance and inherently offers a much higher theoretical upper bound for performance. Despite being the most promising paradigm, realizing the full potential of latent feature transmission is remarkably difficult. Our pilot study in Fig. 1A demonstrates that naively injecting latent features from VLM into the E2E model yields only marginal planning improvements. Consequently, forcing a better alignment between these distinct feature spaces typically demands the intricate architectural modifications seen in recent dual systems (Zhang et al., 2025b; Tang et al., 2025b; Chen et al., 2024c). However, this inevitably exacerbates the exact model coupling we originally sought to eliminate. This dilemma naturally prompts a critical question that *Is designing intricate fusion structures the absolute only way to effectively use cognitive insights from the slow brain to guide the learning of the fast brain?*

Answering this question requires a fundamental understanding of why naive latent feature injection fails in the first place. If the failure stemmed purely from an inadequate architecture, resolving it would inevitably demand the exact complex designs that increase model coupling and violate our original decoupling objective. However, our investigation reveals that the true bottleneck is not a structural deficiency. Instead, the system falls into an underlying optimization trap during cross-model fusion known as *shortcut learning*. Because pre-defined E2E models have already established comfortable optimization trajectories with extremely low regression errors, they instinctively treat newly injected VLM latents as perturbation noise rather than informative priors, driven by both semantic misalignment and substantial information redundancy with the E2E model's native visual features. Under traditional regression training of E2E models, the network rapidly minimizes absolute errors and generates a strong gradient rejection effect, causing the gradients backpropagated to the fusion module to vanish almost entirely (Fig. 1B). Eventually, this causes the fast E2E model to completely ignore the advanced reasoning provided by the slow VLM.

Recognizing the inherent pitfalls of naive feature injection, we argue that effectively harnessing a VLM requires a paradigm shift that replaces dense, unaligned features with concise semantic guidance, while fundamentally rectifying the underlying optimization process. Specifically, we propose

MACA, a novel model-agnostic cognitive plugin. We define "model-agnostic" strictly as E2E-agnostic, meaning MACA maintains a stable upstream VLM as a unified cognitive brain to seamlessly integrate with any downstream E2E driving architecture. First, bypassing high-dimensional feature injection, MACA establishes high-level semantic intentions as a universal communication medium. To keep these intentions precise and lightweight, we introduce a two-stage Sequential Reasoning-to-Decision Training (SRDT) strategy for the VLM. Initially, a high-rank LoRA is trained for comprehensive scene and risk understanding. Then, a low-rank LoRA is fine-tuned to generate compact outputs containing only final decisions and their critical causal triggers. A minimalist network then maps these outputs into continuous latent priors parseable by the planner, requiring no intrusive architectural changes. Second, to break the optimization stagnation caused by gradient vanishing, we propose Relative Advantage Preference Optimization (RAPO). By evaluating a VLM-augmented differentiable E2E model against a frozen pre-trained E2E baseline, RAPO derives instance-level reward and penalty signals from the relative driving performance of the two models. Notably, in industrial practice, this frozen baseline can naturally be a previous-generation production model, effectively repurposing legacy systems to accelerate the iteration of next-generation architectures. This comparison encourages the planner to internalize high-level cognitive priors from the VLM. Ultimately, this approach underpins driving stability in regular scenarios while significantly boosting decision-making success in corner cases, guaranteeing comprehensive outperformance over the corresponding baseline (as shown in Fig. 3).

In summary, our contributions are threefold: ♥ **Insight.** We reveal that "plugin failure" in naive feature injection is an optimization trap, not a structural flaw. Cross-modal shortcut learning and gradient vanishing cause the pre-trained E2E backbone to actively suppress semantic priors as noise. ♥ **Method.** We propose MACA, a model-agnostic cognitive plugin. It extracts precise semantic intentions for non-intrusive injection through two-stage SRDT, and couples this with RAPO to break optimization stagnation and deeply internalize VLM cognition. ♥ **Results.** We demonstrate MACA's generality across diverse SOTA E2E architectures (*e.g.*, DiffusionDrive, SparseDrive, MomAD, GoalFlow). Evaluations confirm it ensures driving stability in regular scenarios while significantly boosting success rates in corner cases, comprehensively outperforming the baselines.

2 Related Work

End-to-End Autonomous Driving. To overcome the inherent information loss and error accumulation of traditional modular architectures, the E2E paradigm maps sensor observations directly to planning outputs (Hu et al., 2023; Chen et al., 2024a; Yang et al., 2025b). Early pioneering works, such as Alvinn (Pomerleau, 1988) and the NVIDIA E2E CNN (Bojarski et al., 2016), adopted a strictly planning-only approach. However, their black-box nature hindered interpretability. Consequently, mainstream research shifted toward multi-task E2E architectures to provide intermediate supervision. Frameworks like the P3-series (Sadat et al., 2020; Hu et al., 2022) and UniAD (Hu et al., 2023) unify perception, prediction, and planning through dense BEV representations. To further improve computational efficiency and structural modeling, this scene representation paradigm evolved into vectorized (Jiang et al., 2023) and sparse approaches (Zhang et al., 2024a; Sun et al., 2024; Hu et al., 2026), as well as recent graph-based (Zhang et al., 2024b) and Gaussian-centric (Zheng et al., 2024b) methods. Concurrently, to address the inherent multimodality of driving behaviors, advancements in trajectory policy have been developed. These include probabilistic modeling (Chen et al., 2024b), multi-target distillation (Li et al., 2024b), online evaluation via BEV world models (Li et al., 2025c; Zheng et al., 2025), and generative strategies (Liao et al., 2025a; Zou et al., 2025; Zheng et al., 2024a; Zhao et al., 2025).

Vision-Language-Action Models and Dual Systems. VLA models attempt to bridge the cognitive gap in autonomous driving by integrating the reasoning capabilities of VLMs. To map multimodal inputs directly to control actions, recent research has improved 3D spatial grounding (Wang et al., 2024; 2026; Li et al., 2025b) and scenario reasoning through mechanisms like Chain-of-Thought (Zhou et al., 2025; Liao et al., 2025b; Liu et al., 2025). While these VLM-native approaches show promise in long-tail scenarios, their high inference latency and training costs hinder real-time deployment. The dual-system architecture, which combines a fast-reactive E2E model with a slow-reasoning VLM, offers a practical way to balance execution efficiency and cognitive depth (Yang et al., 2025b). Within the dual-system paradigm, communication strategies between heterogeneous modules take several forms. Generating explicit waypoints (Tian et al., 2024; Chen et al., 2025)

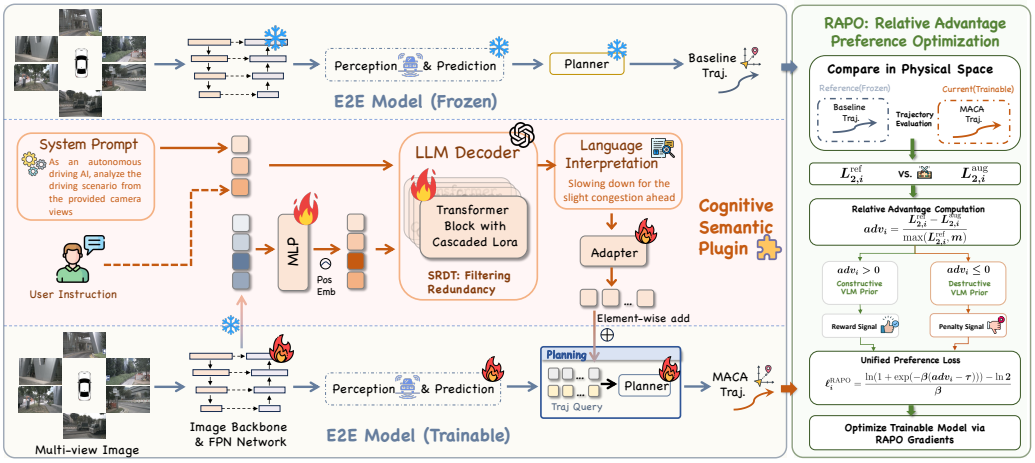


Figure 2: **Overall architecture of MACA**, an E2E-agnostic plugin. An upward interface reuses E2E backbone visual tokens; SRDT produces compact semantic intentions; a downward interface injects them via $\oplus$ into the planner; RAPO optimizes with relative trajectory advantage. Snowflake and fire icons denote frozen and trainable modules.

provides direct planning guidance but can underutilize the language model's semantic reasoning capacity. Mapping decisions to discrete meta-actions using fixed embedding vocabularies (Qian et al., 2024; Jiang et al., 2024; Gariboldi et al., 2025) establishes clear boundaries but often creates an information bottleneck due to the restricted representation space. Injecting latent features, such as the hidden features of the LLM (Zhang et al., 2025b) or the hidden state of the terminal token from the last layer (Chen et al., 2024c; Tang et al., 2025b), preserves denser semantics at the cost of increased architectural coupling required to align distinct modalities. This feature injection approach also encounters the optimization challenge of shortcut learning, where pre-trained E2E models that already achieve low trajectory errors tend to reject new gradients, preventing the fast-reactive planning module from using the injected semantic priors. To address this limitation, MACA bypasses dense high-dimensional feature injection by extracting concise semantic intentions and introduces a preference optimization strategy to mitigate optimization stagnation.

3 Method

3.1 Overview

To address the information bottlenecks of discrete actions and avoid the complex architectural modifications of dense feature alignment, we propose MACA, a model-agnostic cognitive plugin (Fig. 2). We named our framework MACA after the plant known for boosting human energy, as it similarly uses VLMs to enhance virtually any E2E autonomous driving model. Recognizing that naive feature injection fails due to optimization traps rather than structural flaws, MACA avoids complex fusion structures and instead uses high-level semantic intentions as a lightweight communication medium.

The MACA pipeline operates through two non-intrusive interfaces and a unified optimization paradigm. To reduce computational overhead, an upward interface allows the VLM to directly reuse features from the native visual backbone. A downward interface then translates textual decisions into continuous latent priors, which are injected into the E2E planner via simple element-wise addition. To ensure this cognitive guidance translates into performance improvements, MACA employs a two-stage SRDT strategy to extract concise intentions, and introduces RAPO to mitigate the gradient rejection effect of the baseline network.

3.2 Model-Agnostic Architectural Interfaces

To achieve cross-model cognitive fusion without disrupting the internal computational graphs of existing E2E systems, MACA strictly avoids any intrusive architectural modifications (*e.g.*, embedding

complex parameterized fusion modules). Instead, we design two minimalist, unidirectional interfaces, allowing the cognitive plugin to adapt to diverse E2E baselines without redesign. **Upward Interface: Agnostic Visual Bridging.** Standard dual-system approaches often use VLMs with independent image encoders (*e.g.*, CLIP-ViT (Radford et al., 2021)) alongside the E2E system's native visual backbone, creating computational redundancy. MACA omits the dedicated VLM encoder to share computation. This allows direct reuse of features from the pre-trained native backbone without architectural changes, regardless of the visual topology. For E2E systems using ViT, the feature space is natively compatible with mainstream VLMs, supporting joint training. For legacy CNN-based systems such as ResNet (He et al., 2016), compatibility requires a lightweight mapping. Specifically, we extract feature tensors from the deepest layer of the E2E backbone, denoted as $\mathbf{F}_{\text{vis}} \in \mathbb{R}^{B\times N_c\times C\times H\times W}$, where H and W represent the spatial resolution after downsampling. These deep features provide the semantic information necessary for the LLM, while the high downsampling rate maintains a compact spatial grid. To prepare these feature maps for the LLM, they must be converted into a 1D token sequence. However, standard flattening disrupts the inherent 2D topology of the convolutional features. We address this by first aligning the channel dimensions of $\mathbf{F}_{\text{vis}}$ to the LLM's embedding dimension D via an MLP, and then adding a 2D sinusoidal positional encoding $\mathbf{P}_{\text{2D}}$ to maintain spatial indexing within each view. Finally, the enriched feature maps are spatially flattened into a sequence of visual tokens, concatenated with the text prompt tokens $\mathbf{T}_{\text{text}}$, and fed into the LLM for scene reasoning.

$$\mathbf{T}_{\text{dec}} = \text{LLM}\left(\left[\mathbf{T}_{\text{text}}, \text{Flatten}\Big(\text{MLP}(\mathbf{F}_{\text{vis}}) + \mathbf{P}_{\text{2D}}\Big)\right]\right). \tag{1}$$

Downward Interface: Non-Intrusive Semantic Injection. Once the VLM generates the reasoning output $\mathbf{T}_{\text{dec}}$, the downward interface translates it into an implicit prior for the physical planner. To avoid information bottlenecks common in discrete meta-actions, $\mathbf{T}_{\text{dec}}$ acts as more than an isolated control command. Following the SRDT strategy (detailed below), it comprises the final driving behavior and its causal triggers (*e.g.*, "the vehicle should remain stationary until the traffic light turns green"). This transmits high-level semantic intentions while remaining lightweight.

Given the reasoning output $\mathbf{T}_{\text{dec}}$, a pre-trained CLIP text encoder extracts the corresponding semantic feature vector $\mathbf{f}_{\text{text}}$. The inference latency of this textual encoding is negligible (measured at approximately 4 ms), which is offset by the computational savings from discarding the VLM's dedicated visual backbone. To bridge the gap between the VLM's linguistic latent space and the E2E model's driving latent space, we employ a lightweight MLP (composed of fully connected layers and ReLU activations) for cross-model dimensional alignment. This yields a semantic intention latent variable $\mathbf{z} \in \mathbb{R}^{B\times d}$ that matches the feature dimension d of the E2E planner. To maintain compatibility, we avoid operations that alter the baseline's internal attention topology or forward computational flow. MACA uses a simple element-wise addition ($\oplus$) to inject the semantic latent variable as a global condition bias. For query-based architectures like DiffusionDrive (Liao et al., 2025a), the mapped intention is broadcast and added to the planner's initial trajectory query $\mathbf{h}_0$:

$$\mathbf{z} = \text{MLP}_{\text{align}}(\mathbf{f}_{\text{text}}), \qquad \tilde{\mathbf{h}}_0 = \mathbf{h}_0 \oplus \text{Broadcast}(\mathbf{z}) \tag{2}$$

This addition integrates high-level semantic intentions into the trajectory decoding pipeline without modifying native operators, guiding trajectory generation.

3.3 Training and Optimization Strategy

Direct injection of VLM latent features into pre-defined E2E models typically triggers shortcut learning. Since the planning module generates a gradient rejection effect against perturbation during updates, it tends to suppress injected semantic priors. To resolve this stagnation and mitigate vanishing gradients, MACA introduces a unified training and optimization paradigm for cross-modal fusion.

Sequential Reasoning-to-Decision Training. Fine-tuning the VLM presents an optimization conflict between mastering complex reasoning and outputting precise decisions. To address this, we introduce the SRDT strategy: we first train a high-rank LoRA for comprehensive scene and risk understanding, and subsequently apply a low-rank LoRA to produce compact outputs containing only final decisions and causal triggers. This rank-constrained approach filters out verbose descriptions while providing the downstream planner with precise and lightweight semantic intentions.

Relative Advantage Preference Optimization. Standard absolute error metrics (such as the L_1 regression losses commonly used by underlying E2E networks to optimize incremental displacements)

are often insufficient to compel the planner to utilize the injected semantic priors. Because the predefined baseline model has typically converged along a low-error trajectory, it tends to suppress the VLM latents during training, leading to severe vanishing gradients in the semantic fusion module. To break this feature suppression, we do not rely solely on underlying regression constraints and introduce an online preference learning mechanism driven by relative physical driving performance.

Unlike recent LLM preference optimization paradigms such as DPO (Rafailov et al., 2023) and the RL-based GRPO (DeepSeek-AI, 2025), RAPO introduces a fundamentally distinct, non-RL mechanism tailored for cross-modal end-to-end (E2E) driving. The mentioned RL techniques typically operate within discrete semantic spaces and require computationally expensive reinforcement learning loops or extensive response sampling to optimize a single language policy. By completely bypassing the massive training overhead and sample inefficiency inherent to standard RL, RAPO offers a highly streamlined solution for heterogeneous fusion between a VLM and a pre-converged E2E planner. Rather than comparing language outputs or estimating value functions, RAPO grounds its optimization entirely in the continuous physical space. For each driving sample, RAPO compares the trajectory error of the VLM-augmented planner with that of a frozen E2E baseline and uses the performance gap as an instance-level relative reward or penalty signal. It applies stronger correction when the augmented planner underperforms the baseline and reduces the correction once the augmented planner performs better. This asymmetric treatment helps the planner internalize useful VLM priors while suppressing harmful semantic interference.

To formalize these dynamic preference signals, let $L_{2,i}^{\text{ref}}$ and $L_{2,i}^{\text{aug}}$ denote the absolute trajectory displacement errors for the frozen reference baseline and the VLM-augmented model on sample i, respectively. For any instance in the valid sample subset $\mathcal{V}$ (*i.e.*, samples where both branches yield comparable valid trajectories), the relative advantage (adv_i) is formulated as follows:

$$adv_i = \frac{L_{2,i}^{\text{ref}} - L_{2,i}^{\text{aug}}}{\max(L_{2,i}^{\text{ref}}, m)}, \quad i \in \mathcal{V} \tag{3}$$

Here, adv_i measures the relative advantage from the VLM, where the scalar m prevents numerical instability when the baseline error approaches zero. Based on this advantage, the preference loss is formulated as:

$$\ell_i^{\text{RAPO}} = \frac{\text{softplus}(-\beta(adv_i - \tau)) - \ln 2}{\beta}, \quad i \in \mathcal{V} \tag{4}$$

The hyperparameter τ defines the target advantage margin, β modulates gradient sensitivity, and the offset $-\ln 2$ centers the loss function at zero.

This instance-level loss adjusts the optimization strength of each sample according to the relative performance of the augmented and baseline planners. In regular scenarios where the native visual baseline already performs well, injected textual priors may introduce noise, yielding $adv_i \leq 0$. Samples with $adv_i < 0$ are hard for cross-modal fusion because the VLM-augmented planner fails to outperform the baseline. The softplus term assigns these samples a larger positive penalty and stronger corrective gradients, suppressing harmful textual interference and maintaining driving stability. Conversely, when VLM reasoning reduces the trajectory error, $adv_i > 0$. Once the relative advantage exceeds the target margin τ, ℓ_i^{RAPO} becomes negative and serves as a relative reward signal. The corresponding gradient decreases as the advantage grows, avoiding excessive correction of beneficial VLM priors.

Global Loss Construction and Unified Objective. We compute the global average preference loss $\mathcal{L}_{\text{RAPO}}$ over the valid subset $\mathcal{V}$. To align the injected textual semantics with driving trajectory representations, we introduce a trajectory-conditioned semantic alignment loss $\mathcal{L}_{\text{sim}}$. However, to prevent forced global alignment from causing gradient interference and degrading the VLM's logical outputs, we employ an online hard-example mining mechanism. Let the sample-wise cosine similarity loss be $\ell_i^{\text{sim}} = 1 - \cos(\mathbf{z}_i, \mathbf{g}_{\text{gt},i})$, where the target tensor $\mathbf{g}_{\text{gt},i}$ is derived from the ground truth trajectory via a lightweight encoder mapping into the planner's latent space. Thus, $\mathcal{L}_{\text{sim}}$ aligns the text embedding $\mathbf{z}_i$ with the trajectory embedding $\mathbf{g}_{\text{gt},i}$. Utilizing the previously defined relative advantage, we isolate a hard-example subset $\mathcal{H} = \{i \mid adv_i < 0\}$. We call these samples hard examples because the VLM-augmented planner fails to outperform the baseline, indicating difficulty in cross-modal fusion and optimization. The global losses $\mathcal{L}_{\text{RAPO}}$ and $\mathcal{L}_{\text{sim}}$ are computed as follows:

$$\mathcal{L}_{\text{RAPO}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \ell_i^{\text{RAPO}}, \qquad \mathcal{L}_{\text{sim}} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \ell_i^{\text{sim}} \tag{5}$$

This masked normalization strategy ensures that alignment penalties are only triggered when the VLM's spatial priors misguide the physical trajectory. When the VLM provides positive guidance, these constraints are relaxed to protect the learned cross-modal representations.

To prevent un-converged baseline features from generating misleading preference gradients early in training, we introduce an epoch-based warmup coefficient $\lambda_{\text{RAPO}}(t)$ to progressively activate the preference loss. The unified training objective ($\mathcal{L}_{\text{total}}$) comprises the E2E base regression loss, the RAPO loss, and the trajectory-conditioned semantic alignment constraint:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{E2E-base}} + \lambda_{\text{RAPO}}(t)\mathcal{L}_{\text{RAPO}} + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}} \tag{6}$$

Driven by this dynamically adjusted global objective, the planner more effectively internalizes VLM cognitive priors while preserving the continuous-control capability of the baseline.

4 EXPERIMENTS

In this section, we evaluate MACA across several E2E architectures and real-world benchmarks to demonstrate its general applicability and effectiveness in overcoming network inertia. To specifically verify its architecture-independent nature, we integrate MACA into four existing E2E methods without altering their native network topologies. These selected baselines include SparseDrive (Sun et al., 2024) based on entirely sparse queries, DiffusionDrive (Liao et al., 2025a) employing a truncated diffusion policy, MomAD (Song et al., 2025) which enhances planning stability through momentum mechanisms, and GoalFlow (Xing et al., 2025). All evaluation protocols strictly follow the official nuScenes (Caesar et al., 2020) and NAVSIM (Dauner et al., 2024) guidelines.

Implementation Details. We train a Qwen2.5-3B (Team, 2024) LLM on the OmniDrive-nuScenes (Wang et al., 2024) Offline QA subset. SRDT optimizes a dual-MLP projector, then applies LoRA with $r = 128$ and $r = 8$ to compress intentions. During RAPO training, the frozen VLM uses ground-truth semantics, while the planner uses VLM-predicted intentions at inference.

Method	Venue	Input	Backbone	L_2 (m) ↓				Col. Rate (%) ↓			
				1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD (Hu et al., 2023)	CVPR 2023	Camera	ResNet101(He et al., 2016)	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61
VAD (Jiang et al., 2023)	ICCV 2023	Camera	ResNet50(He et al., 2016)	0.41	0.70	1.05	0.72	0.03	0.19	0.43	0.21
LAW (Li et al., 2024a)	ICLR 2025	Camera	-	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30
UncAD (Yang et al., 2025a)	ICRA 2025	Camera	ResNet50(He et al., 2016)	0.29	0.57	0.95	0.60	**0.00**	0.04	0.18	0.07
BridgeAD (Zhang et al., 2025a)	CVPR 2025	Camera	ResNet50(He et al., 2016)	0.29	0.57	0.92	0.59	0.01	0.05	0.22	0.09
VR-Drive (Cho et al., 2025)	NeurIPS 2025	Camera	ResNet50(He et al., 2016)	0.29	0.57	0.95	0.60	0.01	0.03	0.14	0.06
Hip-AD (Tang et al., 2025a)	ICCV 2025	Camera	ResNet50(He et al., 2016)	0.28	0.53	0.87	0.56	0.01	0.05	0.15	0.07
SynAD (Kim et al., 2025)	ICCV 2025	Camera	ResNet50(He et al., 2016)	0.52	0.78	1.10	0.80	0.04	0.10	0.20	0.11
DistillDrive (Yu et al., 2025)	ICCV 2025	Camera	ResNet50(He et al., 2016)	0.28	0.54	0.83	0.57	**0.00**	0.03	0.17	0.06
FlowAD (Guo et al., 2026)	ICLR 2026	Camera	ResNet50(He et al., 2016)	0.28	0.55	0.84	0.56	0.01	0.03	0.15	0.06
MomAD† (Song et al., 2025)	CVPR 2025	Camera	ResNet50(He et al., 2016)	0.30	0.59	0.98	0.62	0.23	0.19	0.36	0.26
MACA$_{\text{MomAD}}$	–	Camera	ResNet50(He et al., 2016)	0.28	0.54	0.89	0.57	**0.00**	0.05	0.20	0.08
SparseDrive-S† (Sun et al., 2024)	ICRA 2025	Camera	ResNet50(He et al., 2016)	0.30	0.58	0.95	0.61	**0.00**	0.05	0.24	0.10
MACA$_{\text{SparseDrive}}$	–	Camera	ResNet50(He et al., 2016)	0.27	0.52	0.86	0.55	0.01	0.04	0.18	0.08
DiffusionDrive† (Liao et al., 2025a)	CVPR 2025	Camera	ResNet50(He et al., 2016)	0.27	0.54	0.90	0.57	0.03	0.05	0.16	0.08
MACA$_{\text{DiffusionDrive}}$	–	Camera	ResNet50(He et al., 2016)	**0.25**	**0.48**	**0.80**	**0.51**	0.01	**0.02**	**0.12**	**0.05**

Table 1: **Results of planning on the nuScenes (Caesar et al., 2020) validation set.** † denotes the results reproduced by our implementation using the official codebase.

4.1 COMPARISON WITH STATE-OF-THE-ART METHODS

Open-loop Evaluation. Tab. 1 compares the open-loop planning performance on the nuScenes dataset. Integrating the MACA plugin improves the performance of three E2E baseline models. Specifically, applying MACA to DiffusionDrive reduces the average L_2 to $0.51\,\text{m}$ and the average collision rate to 0.05%, outperforming existing state-of-the-art methods. Similarly, MACA lowers the average collision rates for both MomAD and SparseDrive to 0.08%, while reducing their average L_2 from $0.62\,\text{m}$ and $0.61\,\text{m}$ to $0.57\,\text{m}$ and $0.55\,\text{m}$, respectively. These improvements are achieved without architectural modifications to the base networks. This indicates that the proposed framework generalizes well across different visual topologies and query representations.

Non-reactive Evaluation. Tab. 2 details the NAVSIM results. Integrating MACA improves the primary PDMS metric for DiffusionDrive (87.5 to 88.2) and, notably, achieves consistent gains

Method	Venue	Input	Backbone	NC ↑	DAC ↑	EP ↑	TTC ↑	Comf. ↑	PDMS ↑
UniAD(Hu et al., 2023)	CVPR 2023	Camera	ResNet-34(He et al., 2016)	97.8	91.9	78.8	92.9	**100.0**	83.4
PARA-Drive(Weng et al., 2024)	CVPR 2024	Camera	ResNet-34(He et al., 2016)	97.9	92.4	79.3	93.0	99.8	84.0
LTF(Prakash et al., 2021)	PAMI 2023	Camera	ResNet-34(He et al., 2016)	97.4	92.8	79.0	92.4	**100**	83.8
Hydra-MDP++(Li et al., 2025a)	–	Camera	ResNet-34(He et al., 2016)	97.6	96.0	80.4	93.1	**100.0**	86.6
Hydra-MDP(Li et al., 2024b)	CVPR 2024	C & L	ResNet-34(He et al., 2016)	98.3	96.0	78.7	**94.6**	**100.0**	86.5
VADv2(Chen et al., 2024b)	ICLR 2026	C & L	ResNet-34(He et al., 2016)	97.2	89.1	76.0	91.6	**100.0**	80.9
GoalFlow(Xing et al., 2025)	CVPR 2025	C & L	ResNet-34(He et al., 2016)	98.3	93.8	79.8	94.3	**100.0**	85.7
ARTEMIS (Feng et al., 2025)	–	C & L	ResNet-34(He et al., 2016)	98.3	95.1	81.4	94.3	**100.0**	87.0
DiffusionDrive†(Liao et al., 2025a)	CVPR 2025	C & L	ResNet-34(He et al., 2016)	98.1	95.9	81.8	94.4	**100.0**	87.5
MACA$_{\text{DiffusionDrive}}$	–	C & L	ResNet-34(He et al., 2016)	**98.4**	**96.3**	**82.4**	**94.6**	**100.0**	**88.2**
GoalFlow†(Xing et al., 2025)	CVPR 2025	C & L	V2-99(Lee & Park, 2020)	98.0	97.8	84.5	93.7	99.9	89.2
MACA$_{\text{GoalFlow}}$	–	C & L	V2-99(Lee & Park, 2020)	**98.1**	**98.5**	**84.6**	**94.3**	**100.0**	**90.0**

Table 2: **Comparison on NAVSIM (Dauner et al., 2024) navtest split with non-reactive metrics.** C & L denotes the use of both camera and LiDAR as sensor inputs. † denotes the results reproduced by our implementation using the official codebase.

#	*Fusion*	$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{RAPO}}$	L_2 (m) ↓				Col. Rate (%) ↓			
				1s	2s	3s	Avg.	1s	2s	3s	Avg.
①				0.27	0.54	0.90	0.57	0.03	0.05	0.16	0.08
②	✓			0.26	0.51	0.84	0.54	0.02	0.03	0.17	0.07
③	✓	✓		0.26	0.52	0.85	0.54	0.01	0.02	0.16	0.07
④			✓	0.26	0.50	0.82	0.53	0.03	0.05	0.15	0.08
⑤	✓		✓	**0.25**	**0.48**	**0.79**	**0.51**	**0.00**	0.02	0.17	0.06
⑥	✓	✓	✓	**0.25**	**0.48**	0.80	**0.51**	0.01	0.02	**0.12**	**0.05**

Table 3: Component-wise ablation.

Architecture	SRDT	L_2 (m) ↓				Col. Rate (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
MomAD	One-Stage	**0.28**	0.56	0.94	0.60	0.02	0.07	0.21	0.10
	Two-Stage	**0.28**	**0.54**	**0.89**	**0.57**	**0.00**	**0.05**	**0.20**	**0.08**
SparseDrive	One-Stage	0.28	0.53	0.87	0.56	**0.01**	**0.03**	0.19	**0.08**
	Two-Stage	**0.27**	**0.52**	**0.86**	**0.55**	**0.01**	0.04	**0.18**	**0.08**
DiffusionDrive	One-Stage	0.27	0.52	0.87	0.55	0.06	0.06	0.15	0.09
	Two-Stage	**0.25**	**0.48**	**0.80**	**0.51**	**0.01**	**0.02**	**0.12**	**0.05**

Table 4: SRDT ablation.

even on the stronger GoalFlow baseline (89.2 to 90.0). Concurrent gains across sub-metrics (*e.g.*, NC, DAC, EP) confirm that VLM semantics guide planners toward safer, more compliant dynamic decisions, rather than merely adjusting short-term waypoints.

4.2 Component-wise Ablation

To investigate the individual contributions of MACA's components, we conduct ablation studies using DiffusionDrive and other representative baselines.

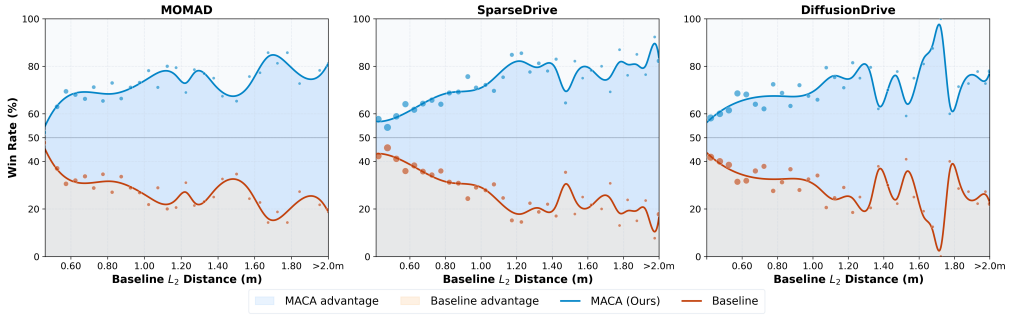


Figure 3: **Win rate of MACA against baselines** rises with scenario difficulty measured by baseline L_2. Point size reflects sample count. Data with $L_2 > 2.0\,\text{m}$ are aggregated.

Mitigating Optimization Stagnation. As shown in Tab. 3, semantic fusion alone (②) reduces the average L_2 from $0.57\,\text{m}$ to $0.54\,\text{m}$, while RAPO without semantic fusion (④) reaches $0.53\,\text{m}$. Combining fusion with RAPO (⑤) further reduces the error to $0.51\,\text{m}$ and the collision rate from 0.08% to 0.06%. These results show that semantic conditioning and the relative objective provide complementary benefits to trajectory optimization.

Tab. 5 compares the DiffusionDrive baseline with a simplified $\text{sgn}(adv_i)$ preference loss and full RAPO. Using only the sign of adv_i lowers the average L_2 to $0.53\,\text{m}$, while RAPO reaches $0.51\,\text{m}$, so the full relative-advantage objective is needed. The lower block of the same table compares injection operators under the same RAPO recipe. Concatenation and cross-attention both reach $0.52\,\text{m}$ average L_2, and element-wise addition reaches $0.51\,\text{m}$. Given the accuracy and the simpler implementation, we keep addition as the default.

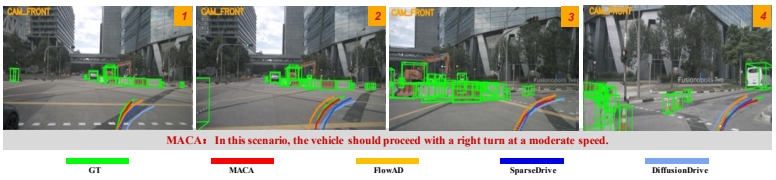


Figure 4: Right-turn trajectory comparison.

Type	Setting	L_2 (m) ↓	Col. Rate (%) ↓
RAPO	Baseline	0.57	0.08
	$\text{sgn}(adv_i)$	0.53	0.18
	RAPO	**0.51**	**0.05**
Injection	Concat	0.52	0.05
	Cross-Attn	0.52	0.05
	Addition	**0.51**	0.05

Table 5: RAPO and injection ablations.

Precise Intention Extraction. Tab. 4 compares single-stage and two-stage SRDT under the same MACA recipe (including RAPO). The single-stage setting uses one rank of LoRA, while the two-stage setting first learns scene understanding with a high-rank LoRA and then compresses outputs with a low-rank LoRA. The two-stage setting lowers average L_2 across baselines (e.g., DiffusionDrive from $0.55\,\text{m}$ to $0.51\,\text{m}$ and MomAD from $0.60\,\text{m}$ to $0.57\,\text{m}$).

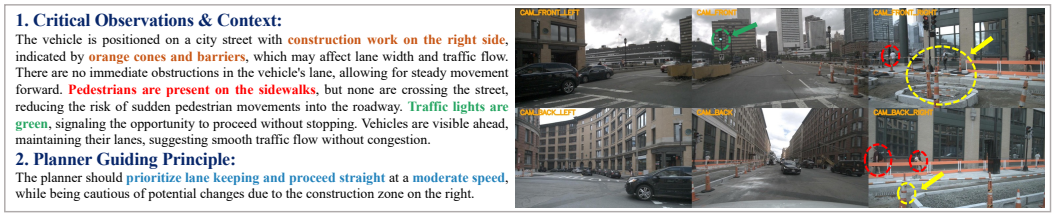


Figure 5: VLM reasoning in a construction scene.

4.3 Further Analysis

Gains in Long-tail Scenarios. Fig. 3 shows the win rate of MACA against the baselines as scenario difficulty varies, with baseline L_2 on the x-axis. As difficulty increases, MACA's win rate grows and the relative improvement becomes more pronounced. This indicates that semantic guidance yields larger gains in long-tail scenes where E2E planning is weaker, and the margin widens with difficulty.

Advantage-Conditioned Alignment. Tab. 6 compares masked and global alignment. Applying $\mathcal{L}_{\text{sim}}$ to all valid samples increases the average L_2 from $0.51\,\text{m}$ to $0.52\,\text{m}$ and the collision rate from 0.05% to 0.07%. RAPO remains active over $\mathcal{V}$, including positive-advantage samples. Only $\mathcal{L}_{\text{sim}}$ is restricted to $\mathcal{H}$ to correct negative transfer while preserving beneficial VLM representations.

Alignment Strategy	Avg. L_2 (m) ↓	Col. Rate (%) ↓
Global ($\mathcal{L}_{\text{sim}}$ on $\mathcal{V}$)	0.52	0.07
Masked ($\mathcal{L}_{\text{sim}}$ on $\mathcal{H}$)	**0.51**	**0.05**

Table 6: Global and advantage-conditioned masked alignment.

Qualitative Analysis. Fig. 4 shows a construction-zone right-turn sequence, where MACA closely tracks the ground-truth trajectory while the baseline planners exhibit larger deviations. In a separate construction scene, Fig. 5 presents the VLM output, which identifies roadside construction, pedestrians, and the green traffic signal and provides a lane-keeping instruction at a moderate speed. Together, they illustrate MACA at both the semantic-reasoning and trajectory-planning levels.

Further ablations, RAPO details, and additional visualizations are in the appendix.

5 Conclusion

In conclusion, we introduced MACA to address the challenge of coupling slow-reasoning Vision Language Models with fast-reactive end-to-end driving architectures. Our investigation shows that naive feature injection failures stem from optimization traps and gradient vanishing rather than architectural flaws. By replacing dense feature injection with precise semantic guidance and a preference optimization strategy, MACA breaks network inertia and helps internalize high-order cognition. Experiments across diverse baseline models support the broad applicability of our approach.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity and readability of the manuscript. We did not use generative AI tools to generate data, develop the method, design experiments, implement the method, or derive theoretical or mathematical claims. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Method formulation, training objectives, and implementation details are provided in the main paper. The appendix reports training procedures, algorithmic details, ablation settings, computational resources, and evaluation protocols. All experiments follow official nuScenes and NAVSIM evaluation settings, and the reproduced baselines are based on official implementations.

REFERENCES

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.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

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*, 2024a.

Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024b.

Xuesong Chen, Linjiang Huang, Tao Ma, Rongyao Fang, Shaoshuai Shi, and Hongsheng Li. Solve: Synergy of language-vision and end-to-end networks for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12068–12077, 2025.

Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. Asynchronous large language model enhanced planner for autonomous driving. In *European Conference on Computer Vision*, pp. 22–38. Springer, 2024c.

Hoonhee Cho, Jae-Young Kang, Giwon Lee, Hyemin Yang, Heejun Park, Seokwoo Jung, and Kuk-Jin Yoon. Vr-drive: Viewpoint-robust end-to-end driving with feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2510.23205*, 2025.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Renju Feng, Ning Xi, Duanfeng Chu, Rukang Wang, Zejian Deng, Anzheng Wang, Liping Lu, Jinxiang Wang, and Yanjun Huang. Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. *arXiv preprint arXiv:2504.19580*, 2025.

Cristian Gariboldi, Hayato Tokida, Ken Kinjo, Yuki Asada, and Alexander Carballo. Vlad: A vlm-augmented autonomous driving framework with hierarchical planning and interpretable decision process. *arXiv preprint arXiv:2507.01284*, 2025.

Mingzhe Guo, Yixiang Yang, Chuanrong Han, Rufeng Zhang, Shirui Li, Ji Wan, and Zhipeng Zhang. FlowAD: Ego-scene interactive modeling for autonomous driving. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=m4JpoJRgAr.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. Springer, 2022.

Weiming Hu, Zhipeng Zhang, Bing Li, Houwen Peng, and Stephen Maybank. Ocean: Object-aware anchor-free tracking with matching-relation learning. *Machine Intelligence Research*, pp. 1–28, 2026.

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. 2023.

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. *arXiv preprint arXiv:2303.12077*, 2023.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

Jongsuk Kim, Jaeyoung Lee, Gyojin Han, Dongjae Lee, Minki Jeong, and Junmo Kim. Synad: Enhancing real-world end-to-end autonomous driving models through synthetic data integration. *arXiv preprint arXiv:2510.24052*, 2025.

Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13906–13915, 2020.

Kailin Li, Zhenxin Li, Shiyi Lan, Jiayi Liu, Yuan Xie, Zuxuan Wu, Zhiding Yu, Jose M. Alvarez, et al. Hydra-mdp++: Advancing end-to-end driving via hydra-distillation with expert-guided decision analysis. *arXiv preprint arXiv:2503.12820*, 2025a.

Tengpeng Li, Hanli Wang, Xianfei Li, Wenlong Liao, Tao He, and Pai Peng. Generative planning with 3d-vision language pre-training for end-to-end autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4950–4958, 2025b.

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*, 2024a.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*, 2025c.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024b.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Haicheng Liao, Hanlin Kong, Bonan Wang, Chengyue Wang, Wang Ye, Zhengbing He, Chengzhong Xu, and Zhenning Li. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*, 2025b.

Xueyi Liu, Zuodong Zhong, Yuxin Guo, Yun-Fu Liu, Zhiguo Su, Qichao Zhang, Junli Wang, Yinfeng Gao, Yupeng Zheng, Qiao Lin, et al. Reasonplan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. *arXiv preprint arXiv:2505.20024*, 2025.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. 1988.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021.

Kangan Qian, Zhikun Ma, Yangfan He, Ziang Luo, Tianyu Shi, Tianze Zhu, Jiayin Li, Jianhui Wang, Ziyu Chen, Xiao He, et al. Fasionad: Fast and slow fusion thinking systems for human-like autonomous driving with adaptive feedback. *arXiv preprint arXiv:2411.18013*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Aditya Ramabadran, Simon Mahns, and Tobias Gessler. Drivingbench: Frontier language models driving a real car through a cone course, 2026. URL `https://drivingbench.com/report`.

Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. Springer, 2020.

Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. *arXiv preprint arXiv:2503.03125*, 2025.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024.

Yingqi Tang, Zhuoran Xu, Zhaotie Meng, and Erkang Cheng. Hip-ad: Hierarchical and multi-granularity planning with deformable attention for autonomous driving in a single decoder. *arXiv preprint arXiv:2503.08612*, 2025a.

Zhipeng Tang, Sha Zhang, Jiajun Deng, Chenjie Wang, Guoliang You, Yuting Huang, Xinrui Lin, and Yanyong Zhang. Vlmplanner: Integrating visual language models with motion planning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5040–5049, 2025b.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, 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.

Jie Wang, Guang Li, Zhijian Huang, Chenxu Dang, Hangjun Ye, Yahong Han, and Long Chen. Vggdrive: Empowering vision-language models with cross-view geometric grounding for autonomous driving. *arXiv preprint arXiv:2602.20794*, 2026.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *CoRR*, 2024.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *CVPR*, 2024.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1602–1611, 2025.

Pengxuan Yang, Yupeng Zheng, Qichao Zhang, Kefei Zhu, Zebin Xing, Qiao Lin, Yun-Fu Liu, Zhiguo Su, and Dongbin Zhao. Uncad: Towards safe end-to-end autonomous driving via online map uncertainty. *arXiv preprint arXiv:2504.12826*, 2025a.

Yixiang Yang, Chuanrong Han, Runhao Mao, Hanshi Wang, Zhiwen Chen, Yantai Yang, Qianli Ma, Xuesong Chen, Shaoshuai Shi, and Zhipeng Zhang. Survey of general end-to-end autonomous driving: A unified perspective. *Authorea Preprints*, 2025b.

Rui Yu, Xianghang Zhang, Runkai Zhao, Huaicheng Yan, and Meng Wang. Distilldrive: End-to-end multi-mode autonomous driving distillation by isomorphic hetero-source planning model. *arXiv preprint arXiv:2508.05402*, 2025.

Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6854–6863, 2025a.

Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024a.

Ruifei Zhang, Junlin Xie, Wei Zhang, Weikai Chen, Xiao Tan, Xiang Wan, and Guanbin Li. Adadrive: Self-adaptive slow-fast system for language-grounded autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5112–5121, 2025b.

Yunpeng Zhang, Deheng Qian, Ding Li, Yifeng Pan, Yong Chen, Zhenbao Liang, Zhiyao Zhang, Shurui Zhang, Hongxu Li, Maolei Fu, et al. Graphad: Interaction scene graph for end-to-end autonomous driving. *arXiv preprint arXiv:2403.19098*, 2024b.

Rui Zhao, Yuze Fan, Ziguo Chen, Fei Gao, and Zhenhai Gao. Diffe2e: Rethinking end-to-end driving with a hybrid action diffusion and supervised policy. *arXiv preprint arXiv:2505.19516*, 2025.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, and Long Chen. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv:2402.11502*, 2024a.

Wenzhao Zheng, Junjie Wu, Yao Zheng, Sicheng Zuo, Zixun Xie, Longchao Yang, Yong Pan, Zhihui Hao, Peng Jia, Xianpeng Lang, et al. Gaussianad: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024b.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. *arXiv preprint arXiv:2507.00603*, 2025.

Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

APPENDIX OVERVIEW

The appendix presents additional algorithmic details, quantitative ablations, qualitative visualizations, and computational-resource details for our Model-Agnostic Contrastive Adaptation (MACA) framework.

1. **Quantitative Ablation on Shortcut Learning (Sec. A)**
   - **Ablation on VLM Signals:** We quantitatively ablate the VLM signals under extreme manipulations (*e.g.*, random noise injection) to expose the underlying shortcut learning and confirm that the baseline merely uses the fusion branch for capacity expansion.
2. **Algorithmic Implementation and Instance-Level Masked Mechanism (Sec. B)**
   - **Algorithmic Implementation of RAPO:** We provide the step-by-step pseudocode for RAPO, detailing the dual-branch forward design and the dynamic preference loss construction.
   - **Ablation on Relative Advantage Distinction:** We ablate the instance-level masked normalization strategy and examine how it corrects hard cases in cross-modal fusion while preserving beneficial VLM representations.
3. **Naive VLM Injection under From-scratch and Masking (Sec. C)**
   - We evaluate naive VLM hidden-feature injection into the E2E planner under from-scratch training and visual-token masking on DiffusionDrive.
4. **Performance under Different Driving Scenarios (Sec. D)**
   - We report scenario-wise planning results (go-straight / turn-left / turn-right) on nuScenes for MACA and the corresponding baselines.
5. **Qualitative Visualizations (Sec. E)**
   - **Robust Planning in Complex Steering Scenarios:** We present visual comparisons demonstrating how MACA utilizes spatial semantic priors to correct lateral steering deviations of pure E2E baselines.
   - **Semantic Adherence and Comprehensive Scene Reasoning:** We visualize MACA's capability to comprehend long-tail semantics (*e.g.*, parked trucks, pedestrians) and adhere to traffic rules (*e.g.*, red lights) via VLM cognitive priors.
6. **Computational Resources (Sec. F)**
   - We detail the hardware specifications (NVIDIA H20 GPUs), training duration, and the default-seed reporting setting.

A QUANTITATIVE ABLATION ON SHORTCUT LEARNING

As highlighted in the Introduction, directly injecting VLM latent features into E2E frameworks frequently triggers *shortcut learning*. During early training, the planning head tends to rely on established visual features, treating newly introduced, unaligned semantic priors as noise. This section examines this failure mode through controlled ablations of the VLM signal.

To verify that the E2E model ignores semantic reasoning rather than facing a capacity bottleneck, we ablate VLM signals on the DiffusionDrive (Liao et al., 2025a) baseline. Tab. 7 presents five configurations: the original baseline, evaluating the trained naive fusion model with the VLM branch disabled at inference ($\emptyset$), replacing real VLM features with random Gaussian noise ($\mathbf{F}_{\text{rand}}$), employing naive feature fusion with actual VLM signals ($\mathbf{F}_{\text{mid}}$), and our MACA.

The results in Tab. 7 further expose the shortcut learning in E2E cross-modal fusion. When the VLM branch is disabled at inference ($\emptyset$), performance degrades severely compared to the original baseline, increasing L_2 to $0.68\,\text{m}$ and collision rate to 0.15%. This degradation suggests that the naive fusion branch ($\mathbf{F}_{\text{mid}}$) contributes to network predictions.

However, when genuine VLM semantic inputs are replaced by random Gaussian noise ($\mathbf{F}_{\text{rand}}$), the system yields an L_2 of $0.56\,\text{m}$ and a collision rate of 0.06%, which is identical to the performance of naive fusion ($\mathbf{F}_{\text{mid}}$) using real language features. Under standard regression metrics such as L_1

Table 7: VLM signal ablation on nuScenes (Caesar et al., 2020) val. $\emptyset$: VLM branch off at inference; $\mathbf{F}_{\text{rand}}$: random features.

Fusion Strategy	Avg. L_2 (m) ↓	Col. Rate (%) ↓
DiffusionDrive (Liao et al., 2025a)	0.57	0.08
DiffusionDrive$_{\text{+Naive}}$ ($\emptyset$)	0.68	0.15
DiffusionDrive$_{\text{+Naive}}$ ($\mathbf{F}_{\text{rand}}$)	0.56	0.06
DiffusionDrive$_{\text{+Naive}}$ ($\mathbf{F}_{\text{mid}}$)	0.56	0.06
DiffusionDrive$_{\text{+MACA}}$ (Ours)	**0.51**	**0.05**

Algorithm 1 Training Pipeline of RAPO

Input: Dataset $\mathcal{D}$, baseline model Θ_{ref} (frozen), augmented model Θ_{aug} (trainable)

Hyperparameters: Margin τ, sensitivity β, clamp m, weights $\lambda_{\text{RAPO}}(t), \lambda_{\text{sim}}$, learning rate η

while model is not converged **do**
  Sample mini-batch $\mathcal{B} \sim \mathcal{D}$ and extract valid sample subset $\mathcal{V} \subseteq \mathcal{B}$
  $\mathbf{P}^{\text{ref}} \leftarrow \Theta_{\text{ref}}(\mathbf{I})$
  $\mathbf{P}^{\text{aug}} \leftarrow \Theta_{\text{aug}}(\mathbf{I}, \mathbf{T})$
  $L_{2,i}^{\text{ref}} \leftarrow \text{L2}(\mathbf{P}_i^{\text{ref}}, \mathbf{G}_i)$ and $L_{2,i}^{\text{aug}} \leftarrow \text{L2}(\mathbf{P}_i^{\text{aug}}, \mathbf{G}_i)$ for all $i \in \mathcal{V}$
  $adv_i \leftarrow \frac{L_{2,i}^{\text{ref}} - L_{2,i}^{\text{aug}}}{\max(L_{2,i}^{\text{ref}}, m)}$ for all $i \in \mathcal{V}$
  $\mathcal{H} \leftarrow \{i \in \mathcal{V} \mid adv_i < 0\}$
  $\ell_i^{\text{RAPO}} \leftarrow \frac{\text{softplus}(-\beta(adv_i - \tau)) - \ln 2}{\beta}$ for all $i \in \mathcal{V}$
  $\mathcal{L}_{\text{RAPO}} \leftarrow \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \ell_i^{\text{RAPO}}$
  if $|\mathcal{H}| > 0$ **then**
    $\mathcal{L}_{\text{sim}} \leftarrow \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left(1 - \cos(\mathbf{z}_i, \mathbf{g}_{\text{gt},i})\right)$
  else
    $\mathcal{L}_{\text{sim}} \leftarrow 0$
  end if
  $\mathcal{L}_{\text{E2E-base}} \leftarrow \text{Reg}(\mathbf{P}^{\text{aug}}, \mathbf{G}) + \text{Cls}(\mathbf{P}^{\text{aug}}, \mathbf{G})$
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{E2E-base}} + \lambda_{\text{RAPO}}(t)\mathcal{L}_{\text{RAPO}} + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}}$
  $\Theta_{\text{aug}} \leftarrow \Theta_{\text{aug}} - \eta \nabla_{\Theta_{\text{aug}}} \mathcal{L}_{\text{total}}$
end while

losses, the E2E model minimizes trajectory errors by using the 1.38M additional parameters in the naive fusion branch as a form of capacity expansion or non-linear regularization. The network ignores the high-order semantic information from the VLM, as substituting it with pure noise causes no performance drop.

This behavior indicates that increasing fusion module complexity cannot overcome the information masking effect of traditional regression objectives. In a low-error converged state, the model maintains visual-dominated trajectories, hindering semantic priors. This explains the performance bottleneck in current VLM fusion methods and highlights the RAPO mechanism's necessity. By optimizing for relative physical driving performance, MACA integrates semantic information, achieving an L_2 of $0.51\,\text{m}$ and a collision rate of 0.05%.

B Algorithmic Implementation and Analysis of the Instance-Level Masked Mechanism

Building on the discussion of RAPO in Section 3.3 of the main manuscript, we describe its algorithmic implementation and evaluate the instance-level mask used for trajectory-conditioned semantic alignment.

B.1 Algorithmic Implementation and Forward Optimization Design

Algorithm 1 formalizes the RAPO training pipeline via a dual-branch forward evaluation mechanism. Specifically, Θ_{ref} denotes the frozen pre-trained E2E baseline model, and Θ_{aug} represents the trainable MACA framework. During each training iteration, both networks process an identical

Setting	Avg. L_2 (m) ↓	Col. Rate (%) ↓
DiffusionDrive baseline	0.57	0.08
Visual-token masking + VLM hidden-feature injection	0.57	0.08
From-scratch training + VLM hidden-feature injection	0.59	0.10

Table 8: **Naive VLM hidden-feature injection checks on DiffusionDrive** (nuScenes val). Both settings inject VLM hidden features into the E2E planner without the full MACA recipe.

batch of multi-view images $\mathbf{I}$. Alongside these visual inputs, Θ_{aug} incorporates the textual semantic intentions $\mathbf{T}$ generated by the language model. Both networks independently output their predicted physical trajectories, denoted as $\mathbf{P}^{\text{ref}}$ and $\mathbf{P}^{\text{aug}}$, which are subsequently evaluated against the ground truth spatial coordinates $\mathbf{G}$ to calculate the respective driving errors. Because the forward pass of Θ_{ref} executes outside the gradient computation graph, this reference evaluation incurs minimal memory and computational overhead, thereby preserving training efficiency.

RAPO Hyperparameters. We set $\tau = 0.05$ and $\beta = 3.0$ for RAPO. The margin τ determines the relative advantage at which the RAPO loss crosses zero, while β controls the sharpness of the transition around this margin.

The instance-level relative advantage adv_i determines the mask used by RAPO. As shown in Algorithm 1, $\mathcal{L}_{\text{RAPO}}$ is evaluated over the valid subset $\mathcal{V}$ and provides an instance-level relative reward or penalty signal. The hard-example subset $\mathcal{H}$ contains samples with $adv_i < 0$, for which the VLM-augmented planner fails to outperform the frozen baseline. These samples are hard from the perspective of cross-modal fusion and optimization. The trajectory-conditioned semantic alignment loss $\mathcal{L}_{\text{sim}}$ is applied only to $\mathcal{H}$. It minimizes the cosine distance between the text embedding $\mathbf{z}_i$ and the trajectory embedding $\mathbf{g}_{\text{gt},i}$ when VLM conditioning degrades trajectory prediction. When the augmented planner matches or outperforms the baseline ($adv_i \geq 0$), this constraint is not applied, avoiding unnecessary regularization of beneficial cross-modal representations.

B.2 Quantitative Ablation on the Relative Advantage Distinction

To evaluate the necessity of applying alignment exclusively to the subset $\mathcal{H}$, we conducted an ablation study. In the Global Alignment setting, we removed the sample distinction and the adv_i-based masked normalization. This configuration applies the $\mathcal{L}_{\text{sim}}$ penalty to all samples in the valid subset $\mathcal{V}$, regardless of whether the VLM reduces the trajectory error.

Tab. 6 compares global and masked alignment. Replacing masked alignment with global alignment increases the average L_2 from $0.51\,\text{m}$ to $0.52\,\text{m}$ and the collision rate from 0.05% to 0.07%. This result exposes two problems with global alignment. For samples with $adv_i < 0$, the VLM-augmented planner underperforms the frozen baseline, indicating negative transfer under semantic conditioning. Restricting $\mathcal{L}_{\text{sim}}$ to these hard examples provides a corrective constraint on the injected representation. For samples with $adv_i > 0$, the augmented planner already outperforms the baseline. Applying $\mathcal{L}_{\text{sim}}$ to these samples forces the text embedding $\mathbf{z}_i$ toward the trajectory embedding $\mathbf{g}_{\text{gt},i}$. This global constraint can compress high-level semantic information into a geometric trajectory representation, weakening reasoning and cross-modal generalization.

Overall, the instance-level mask separates corrective trajectory alignment from beneficial semantic guidance. It constrains samples that exhibit negative transfer while preserving VLM representations that improve planning.

C Naive VLM Injection under From-scratch and Masking

To address whether shortcut learning is limited to adapting a pre-trained planner, and whether a simple forcing strategy can help, we evaluate *naive* VLM hidden-feature injection into the DiffusionDrive planner (without the full MACA recipe) under two settings on nuScenes val.

In the masking setting, we randomly mask VLM visual tokens while still injecting VLM hidden features into the planner. The result matches the DiffusionDrive baseline ($0.57\,\text{m}$ / 0.08%), indicating that simply weakening vision to force VLM reliance does not improve planning. In the from-

Method	Perf. *go straight* ↓ (5309 samples)		Perf. *turn left* ↓ (301 samples)		Perf. *turn right* ↓ (409 samples)		Perf. **Overall** ↓ (6019 samples)	
	Avg. L_2 (m) ↓	Col. Rate (%) ↓	Avg. L_2 (m) ↓	Col. Rate (%) ↓	Avg. L_2 (m) ↓	Col. Rate (%) ↓	Avg. L_2 (m) ↓	Col. Rate (%) ↓
DiffusionDrive (Liao et al., 2025a)	0.55	0.07	0.64	0.12	0.59	0.10	0.57	0.08
DiffusionDrive + **Ours**	**0.51**	**0.05**	**0.57**	**0.08**	**0.52**	**0.03**	**0.51**	**0.05**
SparseDrive-S (Sun et al., 2024)	0.58	0.09	0.75	0.31	0.85	0.08	0.61	0.10
SparseDrive-S + **Ours**	**0.53**	**0.07**	**0.65**	**0.12**	**0.65**	**0.07**	**0.55**	**0.08**
MomAD (Song et al., 2025)	0.61	0.27	0.72	**0.20**	0.68	0.13	0.62	0.26
MomAD + **Ours**	**0.57**	**0.10**	**0.60**	0.32	**0.57**	**0.04**	**0.57**	**0.11**

Table 9: **Performance comparison under different driving scenarios.** MACA reduces L_2 across navigational commands and improves many steering cases.

scratch setting, the E2E planner is trained from random initialization together with the same naive VLM hidden-feature injection, yielding $0.59\,\mathrm{m}$ / 0.10%, which does not surpass the baseline. These checks suggest that neither visual masking nor from-scratch joint training of naive latent injection resolves the optimization issue targeted by MACA.

D Performance under Different Driving Scenarios

Tab. 9 reports planning results under different navigational commands on nuScenes (go-straight / turn-left / turn-right). Baselines are typically stronger in go-straight scenes than in steering. MACA preserves this trend while reducing L_2 in straight driving (*e.g.*, DiffusionDrive from $0.55\,\mathrm{m}$ to $0.51\,\mathrm{m}$). Overall averages remain consistent with the main nuScenes results.

E Qualitative Visualizations

The qualitative evaluations in this section aim to demonstrate the planning robustness of MACA across various complex driving scenarios in the nuScenes dataset, compared to the pure vision baseline and current state-of-the-art methods. Specifically, we illustrate the effectiveness of integrating vision-language model semantic priors from two perspectives: the trajectory planning capability in complex steering scenarios, and the comprehension of complex scenes alongside adherence to traffic rules, which confirms that the introduced semantic priors effectively reflect the actual environmental conditions.

E.1 Planning in Complex Steering Scenarios

Complex steering maneuvers, particularly under conditions like rain, often challenge the planning of pure end-to-end models. Fig. 6 visualizes four challenging driving situations. In the left panel of each scenario, we compare MACA against recent SOTA methods. Guided by high-level semantic instructions such as "proceed with a moderate speed and prepare to make a right turn", MACA generates more reasonable ego-trajectories. The right panels present a direct comparison between MACA and the pure vision baseline, DiffusionDrive (Liao et al., 2025a). While the pure vision baseline frequently exhibits obvious lateral deviations in these situations, MACA effectively corrects these deviations with the assistance of semantic information, yielding trajectories better aligned with the ground truth.

E.2 Semantic Adherence and Scene Reasoning

MACA also demonstrates the ability to understand long-tail obstacles and adhere to traffic rules, which are typically difficult to handle using purely geometric visual features.

As shown in Fig. 7, when the ego-vehicle approaches a red-light intersection, the VLM outputs the semantic understanding to remain stationary until the traffic light turns green. Guided by this semantic instruction, the planner outputs a reasonable stationary trajectory, indicating that the model can effectively understand and respond to the current traffic signal state.

Fig. 8 presents the reasoning results of the model in complex urban scenes. Across different environments, the model accurately identifies key long-tail obstacles, such as parked trucks, traffic cones in construction zones, and pedestrians on the sidewalk. This scene parsing capability provides reliable high-level semantic priors for downstream trajectory planning.

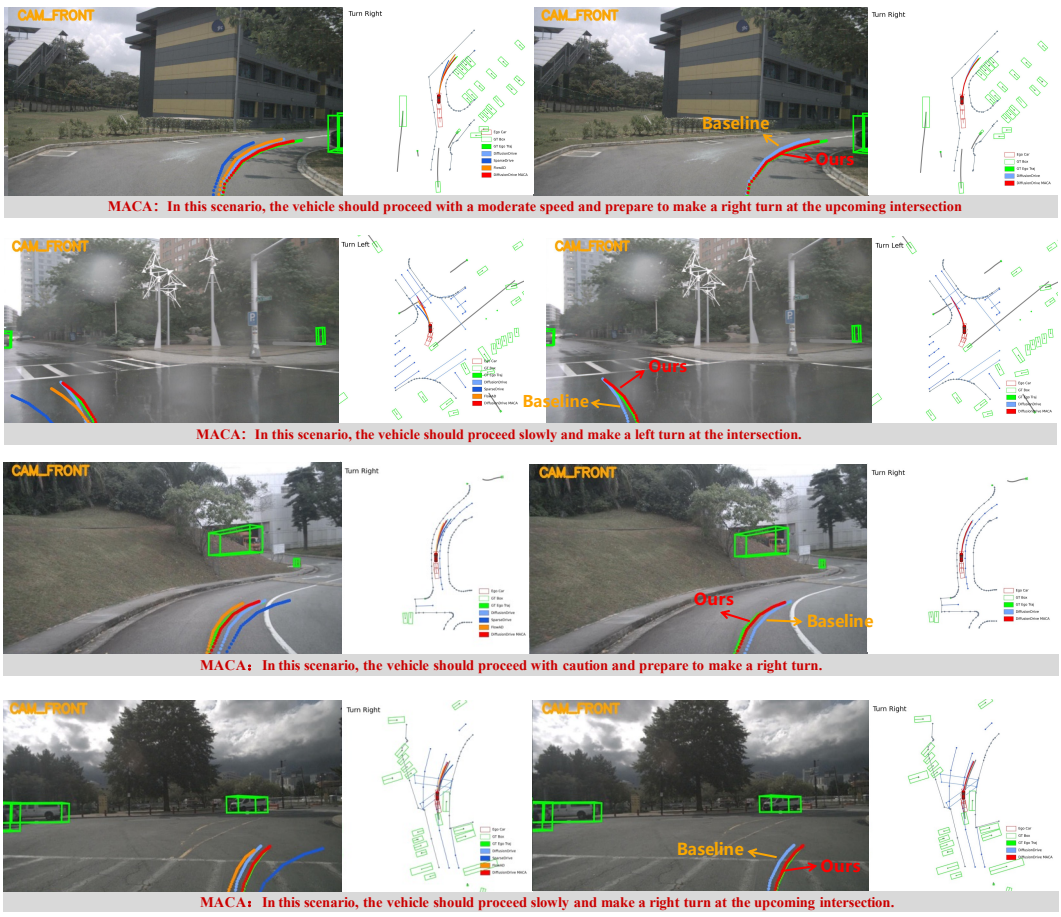


Figure 6: **Qualitative comparison in complex steering scenarios.** The figure visualizes four challenging driving situations. Across all cases, MACA consistently outperforms other SOTA methods (left panels) and corrects the lateral steering deviations of the pure vision baseline (right panels).

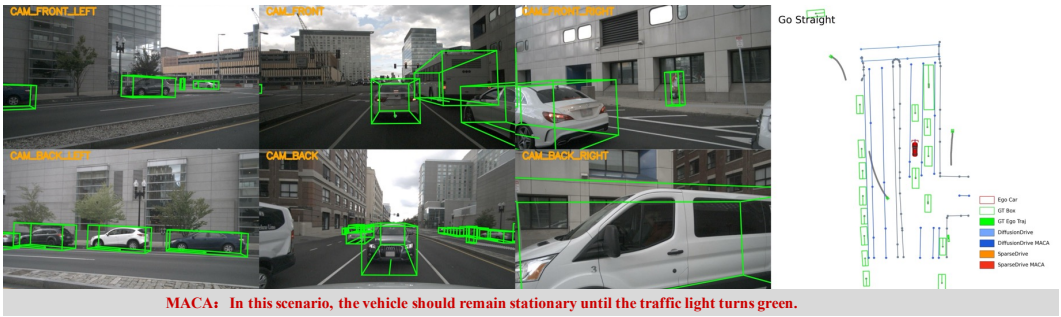


Figure 7: **Visualization of traffic rule adherence.** Guided by the semantic understanding of the red light, MACA stops and waits for the green light.

F Computational Resources

Our experiments were conducted on a compute node with eight NVIDIA H20 GPUs. The training process consists of three stages: approximately one hour for the VLM alignment layer, four hours for the two-stage SRDT rank adaptation, and four hours for the full end-to-end training. A complete training cycle takes around nine hours. We report results from a single run using each baseline's official default random seed (e.g., SparseDrive, MomAD, and DiffusionDrive). To verify

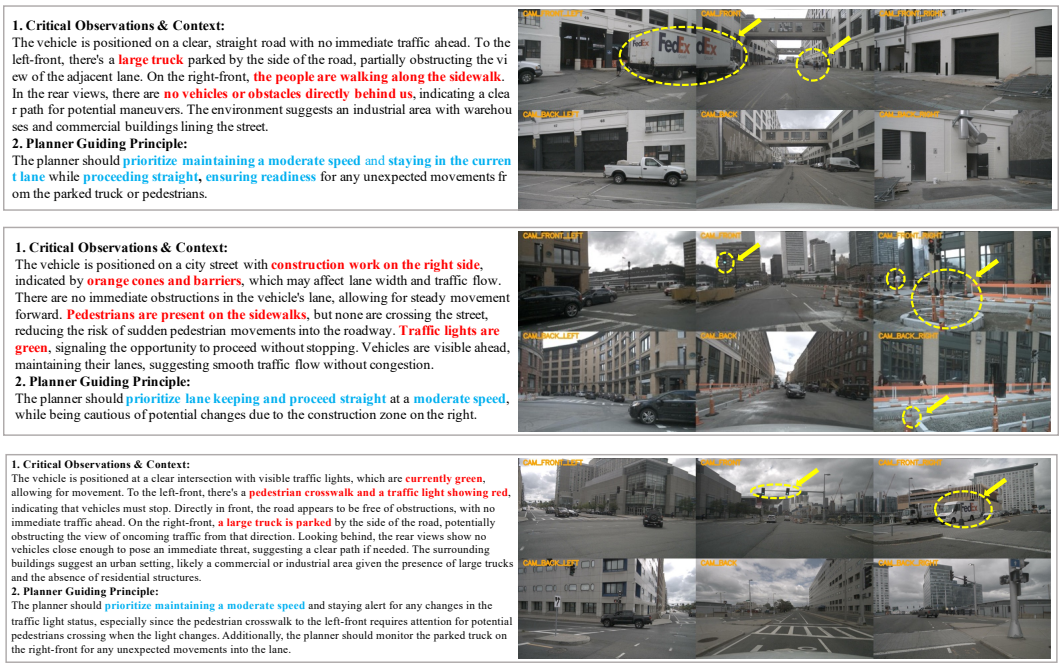


Figure 8: **Comprehensive scene understanding.** The visualizations demonstrate the VLM's reasoning capability to accurately identify critical long-tail obstacles (*e.g.*, parked trucks, construction zones) and deduce high-level guiding principles across complex urban scenarios.

reproducibility, we repeated the full training three times under this same default-seed setting and observed consistent metrics.