DSMOP-Drive: A Deep Reinforcement Learning Framework for Autonomous Driving with Decoupled Safety and Multi-Objective Optimization

1st Zhilin Chen
College of Computer and Data Science
Fuzhou University
Fuzhou, China
202361300@qq.com

2nd Jingtang Chen
Maynooth International College of Engineering
Fuzhou University
Fuzhou, China
jtchen_ai@163.com

3rd Qisong Guo
College of Computer and Data Science
Fuzhou University
Fuzhou, China
1972133773@qq.com

4th Mingjian Fu*
College of Computer and Data Science
Fuzhou University
Fuzhou, China
sinceway@fzu.edu.cn

5th Yuanlong Yu
College of Computer and Data Science
Fuzhou University
Fuzhou, China
yu.yuanlong@fzu.edu.cn

***Abstract*—Deep Reinforcement Learning (DRL) empowers autonomous vehicles with decision-making capabilities to meet diverse driving demands. However, it struggles to jointly optimize multiple conflicting objectives under strict safety constraints. Existing Lagrangian approaches often suffer from gradient conflicts and limited generalization due to coupled representations. To address this, we propose DSMOP-Drive, a primal-based framework decoupling safety rectification from multi-objective optimization. Specifically, it introduces Driving Safety Rectified Multi-Objective Policy Optimization (DSR-MOPO) to harmonize gradients, complemented by a constraint rectification mechanism for immediate safety. Furthermore, we introduce the Dynamic Consistent Joint Embedding Predictive Architecture (DC-JEPA) to disentangle vehicle dynamics from task preferences in latent space, enabling robust adaptation. Experiments in complex highway merge ramp scenarios confirm that DSMOP-Drive achieves a 96.25% success rate under high traffic density, demonstrating low latency in safety response and superior driving utility thanks to the decoupled design.**

***Index Terms*—Autonomous Driving, Safe Reinforcement Learning, Multi-Objective Optimization, Constraint Satisfaction**

I. INTRODUCTION

With the rapid evolution of AI, autonomous driving systems are increasingly deployed in complex, open environments. Deep Reinforcement Learning (DRL) [1] has emerged as a dominant approach for decision and control, outperforming traditional rule-based methods in high-dimensional and nonlinear scenarios. Nevertheless, autonomous driving remains a complex, constrained multi-objective optimization problem [2]. The key challenge is achieving Pareto optimality [3] among these objectives without sacrificing safety.

In response, Safe Reinforcement Learning and Multi-Objective Reinforcement Learning (MORL) [4] have been developed. Although these approaches have made strides, their deployment remains difficult due to structural deficiencies in optimization, safety, and representation. A primary limitation stems from traditional linear scalarization, which frequently masks inherent gradient conflicts [5]. When performance objectives clash with safety constraints, this conflict destabilizes policy updates and prevents the system from converging to Pareto optimality. Compounding this instability is the response latency inherent in Lagrangian relaxation methods [6]. Fundamentally, standard RL architectures suffer from representation entanglement [7]. These models strongly couple objective vehicle dynamics with subjective task preferences, biasing learned features and degrading robustness.

To address these challenges, we propose the DSMOP-Drive framework. Specifically, the Driving Safety Rectified Multi-Objective Policy Optimization (DSR-MOPO) module resolves optimization instability from gradient conflicts by integrating the Natural Conflict Resolution Policy Gradient (NCR-PG). Simultaneously, recognizing the limitations of Lagrangian methods—namely response delays and difficulty in handling sudden hazards—the framework introduces a constraint rectification mechanism. This approach eliminates tuning lag by enforcing boundaries through a highest-priority veto, ensuring instantaneous collision avoidance. Furthermore, to address the representational entanglement between vehicle dynamics and mission preferences, we propose the Dynamic Consistent Joint Embedding Predictive Architecture (DC-JEPA). This architecture significantly enhances model robustness in complex scenarios by separating objective states from subjective rewards through self-supervised auxiliary tasks.

- We propose DSR-MOPO, introducing NCR-PG and a constraint rectification mechanism. This harmonizes conflicting gradients via Fisher regularization and eliminates

Lagrangian response delays via instantaneous constraint rectification.

- We propose DC-JEPA to disentangle objective dynamics from subjective rewards, preventing representation bias and ensuring robust generalization.
- Experiments on highway-env show DSMOP-Drive outperforms state-of-the-art methods, specifically in success rate and average rewards. The code is available at https://github.com/hahahazzza/DSMOP-Drive.

II. RELATED WORK

A. Multi-Objective Reinforcement Learning

MORL is typically studied in two regimes: single-policy learning, which learns one policy balancing multiple objectives [8], and multi-policy learning, which learns a set of non-dominated policies covering diverse trade-offs. In practice, many methods focus on the single-policy setting, employing Multi-Task Learning (MTL) [9] gradient-aggregation strategies to reduce objective interference. These approaches enhance stability by leveraging algorithms or conflict-aware updates like PCGrad [10] and CAGrad [11]. By contrast, the multi-policy regime focuses on approximating the Pareto front to enable a posteriori preference selection. Classic approaches typically learn a parametric manifold whose image in objective space tracks the frontier. Conversely, more recent deep MORL methods utilize hypernetworks to represent the Pareto set, efficiently generating preference-conditioned policies [12].

B. Safe Reinforcement Learning

Safe Reinforcement Learning is typically formulated as a constrained optimization problem [13]. Existing approaches are categorized into: 1) Model-based methods, which approximate safety boundaries via dynamics learning; 2) Safety filters, utilizing formal verification to restrict agents to safe invariant sets; and 3) Constrained policy optimization, imposing constraints on expected cumulative costs, including CRPO [14], PPO-Lag [15], and SAC-Lag [16]. However, existing methods isolate safety constraints from performance objectives, failing to balance them effectively. We propose Safe MORL to ensure safety during exploration within complex traffic flows.

III. METHOD

As shown in Fig. 1, the DSMOP-Drive framework is structured into a three-stage training pipeline. DC-JEPA aligns latent dynamics to extract physical regularities. Then, NCR-PG derives stable updates via Fisher-preconditioned minimax, balancing gradient conflicts under KL constraints. Lastly, a constraint rectification mechanism applies multi-objective updates only if constraints hold, otherwise it switches to Natural Policy Gradients (NPG) [17] to address violations.

A. Driving Safety Rectified Multi-Objective Policy Optimization

While DRL has demonstrated superior potential in high-dimensional decision-making, standard RL inherently struggle to reconcile conflicting multi-objective demands with strict

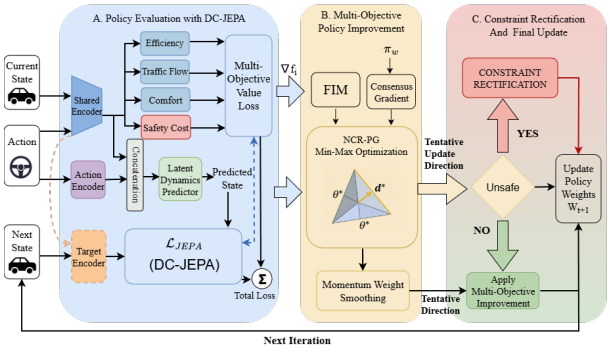


Fig. 1: DSMOP-Drive Framework Overview. (A) Policy Evaluation with DC-JEPA, decoupling dynamics from preferences for robust representation; (B) Multi-Objective Policy Improvement, resolving gradient conflicts via NCR-PG; (C) Constraint Rectification, ensuring strict safety through immediate hard-constraint vetoes.

safety constraints. To overcome these structural limitations, we introduce DSR-MOPO as an advanced primal-based optimization framework. It improves the learning paradigm by structurally decoupling safe Pareto optimality policy learning into three iterative sub-problems: Policy Evaluation, Multi-Objective Policy Improvement, and Constraint Rectification.

Policy Evaluation. We construct a multi-head network $V_\phi(s)$ with a shared encoder $\Phi(s)$ to tackle gradient dominance and conflicts. Four independent heads decouple efficiency, traffic flow, comfort, and safety cost, producing vectorized output that prevents signal masking and supports safety rectification.

$$V(s_t;\phi) = [V_{\text{eff}}(s_t), V_{\text{flow}}(s_t), V_{\text{comf}}(s_t), V_{\text{cost}}(s_t)]^\top. \quad (1)$$

Furthermore, Generalized Advantage Estimation (GAE) is introduced to mitigate variance and oscillation. We construct a stable target $\hat{R}_t^i = \hat{A}_t^i + V_i(s_t;\phi_{\text{targ}})$ using a lagged target value network with parameters ϕ_{targ}, thereby filtering environmental noise. Parameters ϕ are updated by minimizing the Mean Squared Error (MSE):

$$\min_\phi \mathcal{L}(\phi) = \mathbb{E}_\mathcal{D}\left[\sum_i \left(V_i(s_t;\phi) - \hat{R}_t^i\right)^2\right]. \quad (2)$$

This process compels the extractor to balance multi-dimensional features, thereby preventing bias. After K_v iterations, the refined value $\bar{V}_i(s)$ is used to derive NPG, ensuring robust value-based decision-making.

Multi-Objective Policy Improvement. Standard autonomous driving methods using $\nabla_w f_i(\pi_w)$ suffer from Euclidean update oscillation. We mitigate this by optimizing direction d via a Riemannian metric under KL bounds.

$$\max_d \min_{i\in[m]} \xi_i \Delta_i(d) \quad \text{s.t. } D_{\text{KL}}(\pi_w \,\|\, \pi_{w+d}) \le \epsilon_0. \quad (3)$$

Here, $\Delta_i(d) = f_i(\pi_{w+d}) - f_i(\pi_w)$ denotes the performance gain, ξ_i the weight, and ϵ_0 controls update conservativeness.

For tractability, we adopt a local surrogate by applying first- and second-order Taylor approximations around w and replacing the hard KL constraint with a quadratic penalty. This yields the following unconstrained quadratic objective:

$$\max_{d} \min_{i\in[m]} \left\{ \xi_i \nabla_w f_i(\pi_w)^\top d - \frac{\psi_1}{2} d^\top \tilde{F}(w)\, d \right\}. \quad (4)$$

Central to this formulation is $\tilde{F}(w)$ defined as:

$$\tilde{F}(w) = \mathbb{E}_{(s,a)\sim\nu_{\pi_w}} \left[\nabla_w \log \pi_w(a \mid s)\, \nabla_w \log \pi_w(a \mid s)^\top \right]. \quad (5)$$

Although the Fisher information matrix mitigates parameter discontinuity, naive linear aggregation can still cause single-objective dominance or optimization oscillation when conflicting multi-objective gradients arise. Therefore, we introduce conflict regularization to prevent deviation from consensus. The proposed NCR-PG determines the optimal direction by solving the following optimization problem:

$$\begin{aligned} \max_{d} \min_{i\in[m]} \Bigg\{ & \xi_i \nabla_w f_i(\pi_w)^\top d \\ & - \frac{\psi_1}{2} d^\top \tilde{F}(w)\, d - \frac{\psi_2}{2} \|d - v_0\|^2 \Bigg\}. \end{aligned} \quad (6)$$

Here, $\psi_2 \geq 0$ is a hyperparameter limiting deviation between the update direction and the consensus gradient v_0. To solve this minimax problem, we define the composite gradient $\nabla f_\lambda(w) = \sum_{i=1}^m \lambda_i \xi_i \nabla_w f_i(\pi_w)$. Since the objective is concave in d and linear in λ, the minimax theorem allows interchanging the optimization order, yielding the dual form:

$$\min_{\lambda\in S_m} \max_{d} \left\{ \nabla f_\lambda(w)^\top d - \frac{\psi_1}{2} d^\top \tilde{F}(w)\, d - \frac{\psi_2}{2} \|d - v_0\|^2 \right\}. \quad (7)$$

Solving the inner maximization leads to the following optimization problem for λ^*. This formulation seeks an optimal set of weighting coefficients such that the fused gradient direction mitigates conflicts among autonomous driving sub-tasks:

$$\begin{aligned} \lambda^* = \arg\min_{\lambda\in S_m} & \frac{1}{2} (\nabla f_\lambda(w) + \psi_2 v_0)^\top \\ & \cdot (\psi_1 \tilde{F}(w) + \psi_2 I)^{-1} (\nabla f_\lambda(w) + \psi_2 v_0). \end{aligned} \quad (8)$$

The optimal update direction is subsequently given by:

$$d^* = \left(\psi_1 \tilde{F}(w) + \psi_2 I \right)^{-1} \left(\nabla f_{\lambda^*}(w) + \psi_2 v_0 \right). \quad (9)$$

Here, $(\psi_1 \tilde{F}(w) + \psi_2 I)^{-1}$ and $\psi_2 v_0$ provide Riemannian preconditioning and consensus guidance, respectively. These mechanisms prevent policy shifts and gradient conflicts, eliminating oscillations derived from competing objectives.

Constraint Rectification. Following the multi-objective evaluation, the system verifies adherence to hard safety constraints. To overcome the response latency of Lagrangian methods, we implement immediate policy updates using NPG when violations are detected:

$$w_{t+1} = w_t - \eta \tilde{F}(w_t)^\dagger \nabla_w g_{j^*}(\pi_{w_t}). \quad (10)$$

Here, $j^* = \arg\max_j \; g_j(\pi_w)$ denotes the most violated constraint. This mechanism prioritizes safety by forcing the policy into the feasible region, allowing multi-objective updates only when all constraints are satisfied.

Multi-Objective Reward Design and Safety Constraint Construction. To validate DSR-MOPO in highway merge ramp scenarios, we designed a reward function integrating soft and hard constraints. Specifically, the throughput efficiency reward quantifies unit-time capacity and is defined as follows:

$$r_t^{\text{eff}} = \frac{x_t - x_{t-1}}{\Delta t}. \quad (11)$$

Here, x_t, x_{t-1} and Δt denote longitudinal positions and the decision interval. Additionally, a traffic flow following reward smooths gradients by penalizing velocity deviations from surrounding traffic:

$$r_t^{\text{flow}} = - \left| v_{x,t}^{\text{ego}} - \bar{v}_t^{\text{neigh}} \right|. \quad (12)$$

Here, $v_{x,t}^{\text{ego}}$ and $\bar{v}_t^{\text{neigh}}$ denote ego and average neighbor velocities. Additionally, a comfort reward based on acceleration and steering is introduced to eliminate control discontinuities and high-frequency oscillations:

$$r_t^{\text{comf}} = - \left(|a_t| + w_{\text{steer}} \cdot |\delta_t| \right). \quad (13)$$

Here, a_t and δ_t denote the normalized longitudinal acceleration and lateral steering. Furthermore, we define a composite safety cost C_{total}. This metric serves as the trigger for the safety correction module:

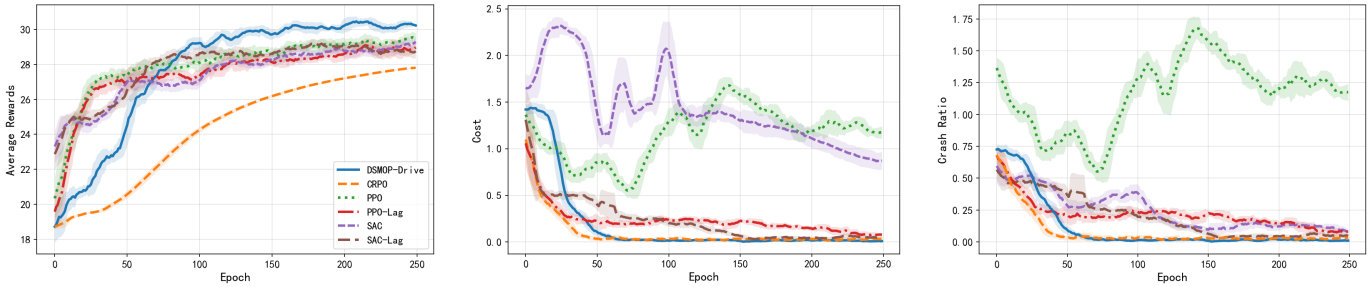
$$\begin{aligned} C_{\text{total}} = & \underbrace{\mathbb{I}(c_{\text{ego}}) + \sum_{v_i \in \mathcal{V}_{\text{others}}} \mathbb{I}(c_i)}_{C_{\text{crash}}} \\ & + \underbrace{\lambda_c \cdot \mathbb{I}(s_t \in \mathcal{S}_{\text{zone}} \wedge a_t = a_{\text{invalid}})}_{C_{\text{action}}}. \end{aligned} \quad (14)$$

Here, $\mathbb{I}(\cdot)$ is the indicator function. The first term C_{crash} aggregates collisions involving the ego vehicle (c_{ego}) and any social vehicle (c_i). The second term C_{action} imposes a penalty λ_c when the agent executes prohibited maneuvers a_{invalid} within restricted zones $\mathcal{S}_{\text{zone}}$. Instead of halting updates, triggering this constraint redirects the optimization to explicitly minimize the cost advantage, allowing for continuous safety recovery without training interruption.

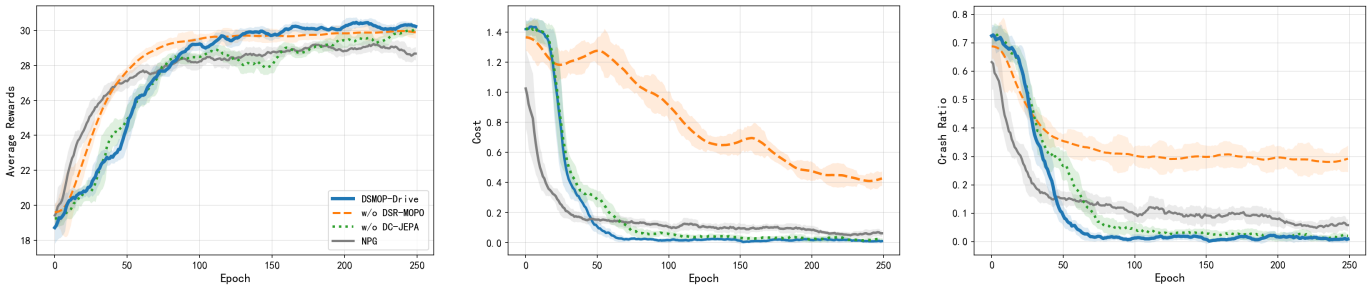
B. Dynamic Consistent Joint Embedding Predictive Architecture

While DSR-MOPO ensures optimization stability, a perception bottleneck persists: scalar rewards entangle objective vehicle dynamics with subjective task preferences. Such entanglement hinders generalization: when task objectives change, the agent must relearn dynamics due to the ambiguity between physics and strategy. We propose DC-JEPA, utilizing auxiliary self-supervised prediction to structurally decouple dynamics from preferences within the feature space.

Action-Conditioned Latent Prediction. To capture task-invariant dynamics, we introduce a latent evolution task. A



(a) Performance comparison with baselines throughout the training process.



(b) Ablation study on the impact of each module during the training phase.

Fig. 2: Comparison of the training process. We present a comparative study with baselines in (a) and an ablation study of key components in (b). The Average Rewards curve shows learning driving utility, the Crash Ratio curve indicates safety convergence, and the Cost curve reflects constraint satisfaction. Shaded areas represent standard deviation.

predictor h infers the next latent state using embeddings from a shared state encoder $z_t = \varphi(s_t)$ and an action encoder $u_t = \varphi_a(a_t)$:

$$\hat{z}_{t+1} = h([z_t, u_t]). \tag{15}$$

Since future states strictly follow Newtonian mechanics governed by kinematics and control—independent of driver intent—the encoder $\varphi(\cdot)$ minimizes prediction error by prioritizing physical features. This effectively filters out irrelevant policy preferences, ensuring the shared representation contains pure dynamics while delegating preference modeling to upper-level value heads. To stabilize training, we employ a target encoder $\varphi^-(\cdot)$ to generate z^-_{t+1}. Its parameters track the online encoder via Exponential Moving Average (EMA) rather than gradient descent.

$$\varphi^- \leftarrow \tau\varphi^- + (1-\tau)\varphi, \quad \tau \in (0,1). \tag{16}$$

This update strategy yields a smooth, consistent regression target, filtering short-term interaction noise and improving long-term robustness of learned dynamics representations.

Self-Supervised Loss and Joint Optimization. Because terminal transitions include non-dynamic effects, DC-JEPA computes self-supervised loss only on non-terminal transitions. The loss is the Euclidean distance between predicted and target representations in normalized space:

$$\mathcal{L}_{\text{JEPA}} = \mathbb{E}\left[(1-d_t)\left\|\frac{\hat{z}_{t+1}}{\|\hat{z}_{t+1}\|_2+\epsilon} - \frac{z^-_{t+1}}{\|z^-_{t+1}\|_2+\epsilon}\right\|_2^2\right]. \tag{17}$$

Here, $d_t \in \{0,1\}$ indicates whether the transition is terminal. Normalization eliminates scale differences. During policy evaluation, this self-supervised loss acts as a regularizer, jointly optimized with multi-head value regression:

$$\mathcal{L} = \mathcal{L}_V + \lambda_{\text{JEPA}}\mathcal{L}_{\text{JEPA}}. \tag{18}$$

Joint optimization anchors the shared representation to physical laws, ensuring stability despite varying reward weights. This decoupling enhances adaptability, requiring only value-head fine-tuning for new preferences. Crucially, it provides NCR-PG with accurate value estimates, mitigating policy bias induced by representation drift.

IV. EXPERIMENT

A. Experimental Setup

We constructed a highway merge ramp scenario in highway-env to verify robustness, where the ego vehicle merges from an on-ramp into main-road traffic. The model was trained for 250 epochs with a learning rate of 1×10^{-3}, and evaluated over 400 episodes for each of the three density levels: Low (0.5–0.7), Medium (0.5–0.9), and High (0.8–1.0). Traffic density is uniformly sampled between the minimum and maximum values, with higher density yielding smaller gaps. To validate performance across distinct safety paradigms, we benchmark against unconstrained (PPO [18], SAC [19]), Lagrangian (PPO-Lag, SAC-Lag), and primal-dual (CRPO) baselines.

B. Performance Comparison

1) Training Analysis: Based on Fig. 2a, we analyze training dynamics to validate multi-objective optimization, constraint

TABLE I: Performance comparison under different traffic densities. We report the Success Rate, Average Cost, and Average Step Reward across three density levels: Low (0.5–0.7), Medium (0.5–0.9), and High (0.8–1.0).

Density	**Metric**	**Method**					
		PPO	SAC	PPO-Lag	SAC-Lag	CRPO	**DSMOP-Drive (Ours)**
Low Density (0.5–0.7)	Success Rate ↑	88.75%	75.50%	94.50%	91.00%	93.75%	**96.75%**
	Average Cost ↓	0.558	1.240	0.251	0.290	0.310	**0.232**
	Average Step Reward ↑	0.666	0.540	0.663	0.654	0.682	**0.689**
Medium Density (0.5–0.9)	Success Rate ↑	84.25%	72.75%	91.00%	87.25%	83.25%	**97.25%**
	Average Cost ↓	0.814	1.529	0.309	0.391	0.549	**0.307**
	Average Step Reward ↑	0.581	0.500	0.590	0.574	0.588	**0.657**
High Density (0.8–1.0)	Success Rate ↑	56.25%	61.25%	85.00%	76.25%	73.25%	**96.25%**
	Average Cost ↓	1.637	2.008	0.518	0.651	0.793	**0.457**
	Average Step Reward ↑	0.262	0.419	0.420	0.431	0.472	**0.638**

TABLE II: Ablation study of DSMOP-Drive under varying traffic densities, comparing the Full model against variants without DC-JEPA ("w/o DC-JEPA") or DSR-MOPO ("w/o DSR-MOPO"), alongside the NPG baseline.

Density	**Metric**	**Method Variants**			
		DSMOP-Drive (Full)	w/o DC-JEPA	w/o DSR-MOPO	Baseline (NPG)
Low Density (0.5–0.7)	Success Rate ↑	**96.75%**	96.25%	87.25%	88.00%
	Average Cost ↓	**0.232**	0.300	0.392	0.365
	Average Step Reward ↑	**0.689**	0.665	0.641	0.578
Medium Density (0.5–0.9)	Success Rate ↑	**97.25%**	95.50%	78.00%	85.25%
	Average Cost ↓	**0.307**	0.412	0.658	0.387
	Average Step Reward ↑	**0.657**	0.609	0.557	0.487
High Density (0.8–1.0)	Success Rate ↑	**96.25%**	95.25%	71.00%	83.00%
	Average Cost ↓	**0.457**	0.561	0.933	0.533
	Average Step Reward ↑	**0.638**	0.535	0.491	0.386

rectification, and representation decoupling. Regarding safety convergence, unconstrained methods (PPO, SAC) prioritize short-term driving utility, resulting in high and fluctuating collision rates. Although Lagrangian methods demonstrate some improvement, they are hindered by lagged dual variable updates, as the system must accumulate violations to adjust multipliers, thereby retarding convergence. In contrast, DSMOP-Drive achieves rapid convergence, attributed to DSR-MOPO. Unlike passive penalty schemes, this mechanism prioritizes safety by projecting the policy back into the safe region, establishing definitive boundaries during early exploration. Furthermore, concerning the driving utility-exploration trade-off, DSMOP-Drive exhibits initially lower rewards (0–50 epochs), reflecting a conservative exploration strategy, yet steadily surpasses all baselines after 100 epochs. Notably, unlike CRPO, DSMOP-Drive leverages smooth gradient conflict handling to identify optimal solutions within the manifold space. Ultimately, DSMOP-Drive resolves driving safety bottlenecks through rational algorithmic mechanisms rather than simplistic speed reduction.

2) Testing Analysis: Table I compares DSMOP-Drive against baselines across three traffic densities, demonstrating its superiority in handling multi-objective conflicts. Regarding unconstrained algorithms (PPO, SAC), significant instability is observed under high density. Success rate is defined as successfully merging and reaching the destination. Specifically, PPO's success rate drops to 56.25%, while SAC incurs a cost of 2.008. This degradation arises because the absence of explicit constraints in single-scalar reward mechanisms. Conversely, while Lagrangian methods (PPO-Lag/SAC-Lag) enhance safety, they sacrifice driving utility. Although PPO-Lag achieves a 91.00% success rate at medium density, its reward is limited to 0.590, inferior to DSMOP-Drive's 0.657. This gap stems from lagged dual variable updates, where delayed multiplier adjustments force the policy to oscillate between aggressive and conservative behaviors, impeding driving utility. Furthermore, CRPO achieves only a 73.25% success rate with a cost of 0.793 under high density. Despite primal-dual updates, the lack of representation decoupling leaves physical properties entangled with task preferences, limiting robustness. Ultimately, DSMOP-Drive achieves an optimal balance. It attains a 96.25% success rate in high density—outperforming PPO-Lag's 85.00%—via effective constraint rectification. Maintaining success rates above 96% across the density spectrum validates DC-JEPA; this minimal variance demonstrates that decoupling dynamics enhances generalization. Moreover, the reduced cost of 0.457 indicates that NCR-PG mitigates interference in the manifold

space, yielding smoother decision-making.

C. Ablation Study

To validate the effectiveness of the core components within DSMOP-Drive, we conducted ablation studies by removing the DC-JEPA module (w/o DC-JEPA) and the DSR-MOPO module (w/o DSR-MOPO), respectively. We compared these variants against the Full model and the base NPG algorithm, implemented as TRPO [20].

1) Training Analysis: As shown in Fig. 2b, the Full model (DSMOP-Drive, blue solid line) steadily reaches a reward of about 30 after 150 epochs. In contrast, w/o DSR-MOPO (orange dashed line) grows faster initially by ignoring constraints but converges lower, while Baseline NPG (gray solid line) stagnates significantly lower. Cost analysis validates DSR-MOPO's impact: the Full method reduces costs to near-zero, whereas w/o DSR-MOPO stagnates at about 0.4, indicating severe safety violations. This proves that reward penalties alone—without NCR-PG and constraint rectification—are ineffective for hazard avoidance. Additionally, the Crash Ratio plot confirms that while w/o DC-JEPA (green dotted line) eventually matches the Full model's near-zero ratio, w/o DSR-MOPO maintains a high 0.3 ratio. Thus, DSR-MOPO ensures strict safety, while DC-JEPA improves rewards and representation stability.

2) Testing Analysis: Table II presents the ablation study results for different variants. The removal of the DSR-MOPO module significantly impacts safety. Without this module, the success rate in high-density scenarios drops from 96.25% to 71.00%, and the average cost increases to 0.933. This degradation occurs because the agent relies only on reward penalties which are insufficient to strictly prevent dangerous maneuvers, leading to frequent safety violations. The DC-JEPA module mainly influences driving efficiency and generalization. Removing this module leads to a clear drop in the average step reward, from 0.638 to 0.535 in high-density settings, while the success rate remains relatively stable. This reduction in efficiency occurs because the vehicle dynamics and task preferences are not decoupled in the latent space. As a result, the agent fails to learn a generalized physical representation and tends to adopt conservative driving behaviors to maintain safety in complex traffic, thereby sacrificing traffic throughput. Ultimately, the Full model combines DSR-MOPO to guarantee a safety baseline through constraints and DC-JEPA to enhance utility through robust representation learning.

V. CONCLUSION

We propose DSMOP-Drive to address structural limitations in optimization, safety, and representation. DSR-MOPO integrates NCR-PG to balance gradient conflicts via the Fisher information matrix for stable Pareto convergence, while its constraint rectification eliminates Lagrangian latency through priority vetoes, ensuring immediate safety. Additionally, DC-JEPA decouples dynamics from preferences using self-supervision, enhancing generalization. However, this reliance on latent prediction demands a high-quality encoder to filter task-irrelevant features. Experiments confirm DSMOP-Drive outperforms baselines with a 96.25% success rate, excelling in both driving utility and stability. Future work could extend DC-JEPA into a more forward-looking world model to further enhance long-term planning.

REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] H. Wang, Z. Wang, and X. Cui, "Multi-objective optimization based deep reinforcement learning for autonomous driving policy," *Journal of Physics: Conference Series*, vol. 1861, no. 1, p. 012097, 2021.

[3] F. Yang, H. Huang, W. Shi, Y. Ma, Y. Feng, G. Cheng *et al.*, "Pmdrl: Pareto-front-based multi-objective deep reinforcement learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 9, pp. 12 663–12 672, 2022.

[4] K. Van Moffaert and A. Nowé, "Multi-objective reinforcement learning using sets of pareto dominating policies," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3483–3512, 2014.

[5] A. Annunziata, M. Lapucci, P. Mansueto, and D. Pucci, "On the computation of the efficient frontier in advanced sparse portfolio optimization," *4OR – A Quarterly Journal of Operations Research*, 2025.

[6] S. Bessa, D. Dabert, M. Bourgeat, L.-M. Rousseau, and Q. Cappart, "Learning valid dual bounds in constraint programming: Boosted Lagrangian decomposition with self-supervised learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, pp. 11 113–11 121.

[7] Y. Kang, D. Shi, J. Liu, L. He, and D. Wang, "Beyond reward: Offline preference-guided policy optimization," in *International Conference on Machine Learning*, 2023, pp. 15 753–15 768.

[8] S. Parisi, M. Pirotta, and J. Peters, "Manifold-based multi-objective policy search with sample reuse," *Neurocomputing*, vol. 263, pp. 3–14, 2017.

[9] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Neural Information Processing Systems*, 2018.

[10] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Neural Information Processing Systems*, vol. 33, 2020, pp. 5824–5836.

[11] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, "Conflict-averse gradient descent for multi-task learning," in *Neural Information Processing Systems*, vol. 34, 2021, pp. 18 878–18 890.

[12] T. Shu, K. Shang, C. Gong, Y. Nan, and H. Ishibuchi, "Learning pareto set for multi-objective continuous robot control," in *International Joint Conference on Artificial Intelligence*, 2024, pp. 4920–4928.

[13] S. H. Huang, A. Abdolmaleki, G. Vezzani, P. Brakel, D. J. Mankowitz, M. Neunert *et al.*, "A constrained multi-objective reinforcement learning framework," in *Proceedings of the Conference on Robot Learning*, 2021, pp. 883–893.

[14] T. Xu, Y. Liang, and G. Lan, "Crpo: A new approach for safe reinforcement learning with convergence guarantee," in *International Conference on Machine Learning*, 2021, pp. 11 480–11 491.

[15] A. K. Jayant and S. Bhatnagar, "Model-based safe deep reinforcement learning via a constrained proximal policy optimization algorithm," in *Neural Information Processing Systems*, vol. 35, 2022, pp. 24 432–24 445.

[16] X. Yang, H. He, Z. Wei, R. Wang, K. Xu, and D. Zhang, "Enabling safety-enhanced fast charging of electric vehicles via soft actor critic-lagrange drl algorithm in a cyber-physical system," *Applied Energy*, vol. 329, p. 120272, 2023.

[17] S. M. Kakade, "A natural policy gradient," in *Neural Information Processing Systems*, 2001.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*, vol. 80, 2018, pp. 1861–1870.

[20] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, vol. 37, 2015, pp. 1889–1897.