Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 251-280. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1784

Proximal Policy Optimization in Autonomous Driving: A Systematic Review of Methods, Imitation Learning, and Evaluation Practices

Anas Bayu Kusuma, Yogiek Indra Kurniawan, Vektor Dewanto, Wisnu Jatmiko

Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia

Email:anas.bayu@ui.ac.id, yogiek.indra@ui.ac.id, vektor@cs.ui.ac.id, wisnuj@cs.ui.ac.id

Abstract

The development of robust decision-making policies for road-based autonomous vehicles (AV) remains a critical research challenge. While Reinforcement Learning (RL) and Imitation Learning (IL) show promise, the research landscape remains fragmented, particularly regarding Proximal Policy Optimization (PPO). This paper presents a systematic literature review synthesizing PPO applications in autonomous driving. Following the Kitchenham methodology and SEGRESS guidelines, we searched major digital libraries for studies published between 2021 and 2025. From 108 initial records, a rigorous selection process yielded 26 primary studies. Our analysis reveals that standard PPO dominates (73.1%), with modified variants accounting for 23.1%. CARLA serves as the primary evaluation platform (50.0%), with urban driving (42.3%) and lane changing (30.8%) being the most common tasks. IL integration employed diverse approaches including Behavioral Cloning, GAIL, and AIRL. A significant finding is evaluation fragmentation, with 14 unique metrics identified, though collision rate (57.7%) and success rate (30.8%) were most prevalent. Quality assessment showed 46.2% of studies achieved high methodological quality, while transparency emerged as the weakest criterion. Our findings underscore the need for standardized benchmarks, sim-to-real transfer methods, and improved reporting transparency.

Keywords: *Autonomous vehicles, imitation learning, proximal policy optimization, reinforcement learning, systematic literature review*

1. Introduction

The field of autonomous vehicles (AV) has become a prominent area of research, driven by the transformative potential of this technology to improve safety, enhance transportation efficiency, and increase mobility. This review specifically focuses on **road-based autonomous vehicles** such as passenger cars, trucks, and commercial vehicles operating on public roads, and excludes other autonomous systems such as unmanned aerial vehicles (UAVs), unmanned surface vehicles (USVs), and autonomous robotic systems.

The Society of Automotive Engineers (SAE) provides a widely adopted framework classifying driving automation from Level 0 (no automation) to Level 5 (full automation) [1]. Currently, most commercially deployed systems, such as Tesla Autopilot, operate at Level 2 or 3 [2], highlighting the considerable challenges that remain in creating a robust driving policy capable of navigating complex and unpredictable real-world scenarios.

To overcome the limitations of traditional rule-based systems, research has increasingly shifted towards learning-based methods. **Reinforcement Learning (RL)**, which trains agents through environmental interaction [3], has emerged as a powerful paradigm for developing driving policies. The RL algorithm landscape for autonomous driving encompasses several major families, each with distinct characteristics and trade-offs:

Value-based methods such as Deep Q-Networks (DQN) [4] and Rainbow [5] learn action-value functions and have shown success in discrete action spaces. However, their extension to the continuous control problems characteristic of vehicle steering and throttle control remains challenging. *Policy gradient methods* including Advantage Actor-

Critic (A3C) [6], Trust Region Policy Optimization (TRPO) [7], and Proximal Policy Optimization (PPO) [8] directly optimize parametrized policies and naturally handle continuous action spaces. *Actor-critic methods* such as Deep Deterministic Policy Gradient (DDPG) [9], Twin Delayed DDPG (TD3) [10], and Soft Actor-Critic (SAC) [11] combine value function approximation with policy optimization, offering improved sample efficiency but often requiring careful hyperparameter tuning for stable convergence.

Additionally, *Imitation Learning (IL)*, which learns from expert demonstrations [12], provides an alternative or complementary approach. Techniques such as Behavioral Cloning (BC), Dataset Aggregation (DAgger) [13], and Generative Adversarial Imitation Learning (GAIL) [14] address the cold-start problem and can accelerate policy learning when combined with RL methods.

Table 1 compares key characteristics of prominent RL algorithms applied to autonomous driving, highlighting their relative strengths and limitations.

Within this diverse algorithmic landscape, **Proximal Policy Optimization (PPO)** has gained particular distinction in autonomous driving research due to its favorable balance between sample efficiency, implementation simplicity, and training stability. PPO's clipped surrogate objective provides robust policy updates that prevent catastrophic performance degradation, a critical property for safety-sensitive applications where unpredictable policy changes could lead to dangerous driving behavior. Unlike TRPO, which requires computationally expensive second-order optimization, PPO achieves similar stability guarantees using only first-order methods, making it more practical for the high-dimensional state and action spaces characteristic of autonomous driving. Compared to DDPG and TD3, PPO exhibits more consistent convergence across different environments without requiring extensive hyperparameter search, a crucial advantage when transferring policies across varied driving scenarios (urban, highway, adverse weather). Furthermore, PPO's on-policy nature aligns well with safety constraints, as it learns directly from the current policy's experience rather than relying on off-policy data that may reflect outdated or unsafe behaviors.

The algorithm's widespread adoption in autonomous driving research and its unique position balancing stability, efficiency, and implementation accessibility warrant dedicated systematic examination. However, while there is a substantial body of primary research applying PPO to AV, the literature remains fragmented, making it difficult to form a holistic understanding of how PPO is being applied, how it relates to IL techniques, and what challenges persist.

Given the rapid and diverse advancements in this domain, a systematic literature review is required to structure the existing evidence. While broad surveys on deep learning for AV [15], RL for autonomous driving [16], and trajectory prediction [17] exist, none provide a focused synthesis of how PPO, one of the most widely adopted policy gradient algorithms, is specifically being applied to autonomous driving tasks. Furthermore, since PPO-based approaches are often combined with or compared against IL techniques, understanding this relationship is essential. This review aims to fill that gap by providing a systematic, evidence-based analysis of PPO applications in AV, the IL techniques used alongside them, and the evaluation practices employed in this research area.

To achieve this, our systematic literature review is guided by the following research questions:

1) RQ1: How is Proximal Policy Optimization (PPO) being applied within RL frameworks for autonomous driving tasks?
2) RQ2: In the context of these PPO studies, how are IL techniques integrated with or compared to PPO?
3) RQ3: What types of autonomous driving tasks and environments are being used to evaluate these PPO-based and IL policies?
4) RQ4: What performance metrics are used to assess the effectiveness of these driving policies in autonomous driving scenarios?

We focus on studies published between January 2021 and July 2025 to capture the most recent advancements following PPO's maturation as a standard algorithm. We exclude multi-agent reinforcement learning (MARL) studies to maintain a focused scope on single-agent driving policies, as MARL introduces additional complexity warranting separate analysis.

The remainder of this paper is organized as follows. Section II reviews related secondary studies to situate our work and identify the specific gap this review addresses. Section III details the systematic methodology used for study selection and data synthesis. Section IV presents the synthesized results, structured according to the research questions. Section V discusses the implications of our findings and acknowledges the study's limitations. Section VI concludes the paper with a summary of contributions and directions for future research.

This systematic literature review makes the following contributions:

Table 1. Comparison of reinforcement learning algorithms for autonomous driving.

Algorithm	Sample Efficiency	Training Stability	Continuous Control	Implementation Complexity	Key Limitation for AV
DQN	Moderate	Moderate	Poor (discrete only)	Low	Requires action discretization
A3C	Low	Moderate	Good	Moderate	High variance, many parallel workers needed
DDPG	High	Low	Excellent	Moderate	Sensitive to hyperparameters, brittle
TD3	High	Moderate	Excellent	Moderate	Careful tuning required
SAC	High	High	Excellent	High	Computational overhead, entropy tuning
TRPO	Moderate	High	Good	High	Computationally expensive
PPO	**Moderate-High**	**High**	**Good**	**Low**	**Moderate sample complexity**

1) Systematic identification and quality assessment of 26 primary studies (2021-2025) applying PPO to autonomous vehicles
2) Detailed categorization of PPO variants (standard, modified, baseline) with usage patterns and implementation approaches
3) Comprehensive mapping of IL integration techniques with PPO-based systems, including BC, GAIL, AIRL, and hybrid approaches
4) Systematic documentation of evaluation environments, driving tasks, and performance metrics, revealing significant fragmentation (14 unique metrics)
5) Evidence-based identification of research gaps, challenges, and future directions specific to PPO in autonomous driving.

2. Related Work

The application of deep reinforcement learning (DRL) to autonomous driving has gathered substantial research interest, resulting in several comprehensive surveys that synthesize this body of work. This section reviews three influential surveys that inform the current study and identifies the specific gap that motivates this systematic literature review.

2.1. Proximal Policy Optimization: Technical Overview

To provide context for the systematic review findings, this section presents a technical overview of Proximal Policy Optimization (PPO), the algorithm at the center of this study. While readers familiar with PPO may skip this section, we include it to ensure accessibility for those new to the algorithm or seeking a refresher on its key principles.

2.1.1. Policy Gradient Methods and the Trust Region Problem. Proximal Policy Optimization belongs to the family of *policy gradient methods*, which directly optimize a parametrized policy $\pi_\theta(a|s)$ that maps states to actions. Unlike value-based methods (e.g., DQN) that learn action-value functions, policy gradient methods use gradient ascent to maximize expected cumulative reward:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{1}$$

where τ represents a trajectory, γ is the discount factor, and r_t is the reward at time t.

A fundamental challenge in policy gradient optimization is ensuring that policy updates are neither too large (causing performance collapse) nor too small (slowing learning). Trust Region Policy Optimization (TRPO) [7] addressed this by constraining updates to lie within a trust region, ensuring monotonic improvement. However, TRPO's reliance on second-order optimization (computing the Fisher information matrix and conjugate gradient) makes it computationally expensive, particularly for the high-dimensional neural networks used in autonomous driving.

2.1.2. The PPO Clipped Surrogate Objective. Proximal Policy Optimization [8] achieves TRPO's stability guarantees using a simpler, first-order approach. PPO introduces a clipped surrogate objective function that penalizes policy updates that deviate too far from the current policy:

$$L^{CLIP}(\theta) = \mathbb{E}_t\Big[\min\big(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big] \quad (2)$$

where:

- $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio
- $\hat{A}_t$ is the advantage function estimating how much better action a_t is compared to the average
- ϵ is the clipping parameter (typically 0.1 to 0.2)

2.1.3. Intuitive Explanation of the Clipping Mechanism. The clipped objective in Equation 2 implements a conservative update strategy through three cases:

Case 1: Positive advantage ($\hat{A}_t > 0$). If an action is better than average, we want to increase its probability. However, if $r_t(\theta) > 1+\epsilon$, the clipping prevents the probability from increasing too much. The objective becomes $(1+\epsilon)\hat{A}_t$, removing the incentive for further increases.

Case 2: Negative advantage ($\hat{A}_t < 0$). If an action is worse than average, we want to decrease its probability. However, if $r_t(\theta) < 1-\epsilon$, the clipping prevents the probability from decreasing too much. The objective becomes $(1-\epsilon)\hat{A}_t$, removing the incentive for further decreases.

Case 3: Within trust region ($1-\epsilon \leq r_t(\theta) \leq 1+\epsilon$). The clipping has no effect, and standard policy gradient updates apply.

2.1.4. Implementation and Advantages. In practice, PPO is implemented using the following procedure:

1) Collect trajectories by running the current policy $\pi_{\theta_{\text{old}}}$ in the environment
2) Compute advantages $\hat{A}_t$ using Generalized Advantage Estimation (GAE) [18]
3) Optimize the clipped objective (Equation 2) using minibatch gradient ascent for multiple epochs
4) Update the policy: $\theta_{\text{old}} \leftarrow \theta$
5) Repeat

This approach offers several advantages for autonomous driving applications:

- **Training Stability:** The clipping mechanism prevents large, destabilizing policy updates, critical for safety-sensitive domains where erratic behavior could lead to collisions.
- **Sample Efficiency:** By reusing collected trajectories for multiple gradient updates (via the clipping constraint), PPO achieves better sample efficiency than vanilla policy gradients while remaining on-policy.
- **Implementation Simplicity:** Unlike TRPO, PPO requires only first-order gradients and standard optimizers (e.g., Adam), making it easier to implement and scale to large neural networks.
- **Continuous Action Spaces:** PPO naturally handles the continuous control required for vehicle steering angles and throttle/brake commands without action discretization.

2.1.5. Relevance to Autonomous Driving. The properties of PPO align particularly well with the challenges of autonomous driving:

Safety through stable updates. The clipped objective ensures that the driving policy does not suddenly adopt dangerous behaviors, even when learning from exploratory or suboptimal experiences. This is crucial when training in simulators before real-world deployment.

Handling sparse rewards. Autonomous driving tasks often have sparse reward signals (e.g., collision penalties, goal completion rewards). PPO's advantage estimation and multi-epoch updates help extract meaningful learning signals from limited feedback.

Scalability to complex scenarios. PPO's computational efficiency enables training on high-dimensional sensory inputs (camera images, LiDAR point clouds) and complex urban environments with many interacting agents.

These properties have contributed to PPO's widespread adoption in autonomous driving research, motivating this systematic review to synthesize how the algorithm is being applied, adapted, and evaluated across diverse driving scenarios.

2.2. Existing Surveys on DRL for Autonomous Driving

Kiran *et al.* [19] present a comprehensive survey on DRL for autonomous driving, offering an extensive overview of the field's landscape. Their work systematically reviews core RL concepts, including Markov Decision Processes (MDPs), value-based methods such as Deep Q-Networks (DQN), and policy gradient approaches including Proximal Policy Optimization. The survey establishes a useful taxonomy that categorizes the literature according to simulator platforms (CARLA, TORCS, SUMO), application scenarios (lane keeping, intersection navigation, highway driving), and reward function designs. While this survey provides valuable foundational coverage of DRL algorithms for autonomous

vehicles, its broad scope across all DRL methods results in limited depth on any single algorithm. Proximal Policy Optimization receives only brief treatment alongside numerous other algorithms, without detailed analysis of its specific variants, implementation patterns, or comparative advantages in autonomous driving contexts.

Zhu and Zhao [20] contribute a survey that uniquely addresses both DRL and deep imitation learning (DIL) for autonomous driving policy learning. Their work introduces a five-mode taxonomy that classifies how DRL/DIL models integrate into autonomous driving system architectures: (1) end-to-end control, (2) control with high-level commands, (3) motion planning integration, (4) behavior planning integration, and (5) hierarchical planning and control. Beyond architectural considerations, the survey comprehensively reviews task-driven formulations covering state space design, action space design, and reward function engineering. The authors also examine problem-driven methods addressing driving safety, interaction with traffic participants, and environmental uncertainty. However, their extensive coverage across all DRL and DIL algorithms necessarily limits algorithm-specific depth. PPO appears in only two studies within their analysis, leaving substantial gaps in understanding how this increasingly popular algorithm performs across different autonomous driving applications.

Zhao *et al.* [21] provide the most recent survey, proposing a six-mode taxonomy that extends prior classification schemes to accommodate emerging architectural patterns. Their taxonomy distinguishes: (1) end-to-end control, (2) conditional end-to-end, (3) motion planning, (4) behavior planning, (5) hierarchical integration, and (6) multi-agent coordination. The survey notably addresses recent algorithmic developments including constrained policy optimization variants (PCPO, CPO) and multi-agent extensions (MAPPO, QMIX). While this survey captures the field's evolution through 2023, its comprehensive scope across all DRL algorithms means that PPO-specific analysis remains limited. The survey identifies PPO variants but does not systematically examine implementation patterns, integration approaches with imitation learning, or performance characteristics specific to PPO-based autonomous driving systems.

2.3. Research Gap and Motivation

Table 2 summarizes the key characteristics of existing surveys and positions the current study. While the reviewed surveys provide valuable contributions to understanding DRL applications in autonomous driving, a notable gap exists in the literature: no existing work provides systematic, in-depth analysis specifically focused on Proximal Policy Optimization. This gap is significant for several reasons.

First, PPO has emerged as one of the most widely adopted algorithms in autonomous driving research due to its favorable balance between sample efficiency, implementation simplicity, and training stability [8]. The algorithm's clipped surrogate objective provides robust policy updates without the computational overhead of trust region methods like TRPO, making it particularly suitable for the high-dimensional state and action spaces characteristic of autonomous driving tasks. Understanding how researchers adapt, modify, and apply PPO specifically warrants dedicated systematic examination.

Second, the intersection between PPO and imitation learning represents an increasingly important research direction that existing surveys do not adequately address. Approaches such as combining PPO with behavioral cloning for policy initialization, using Generative Adversarial Imitation Learning (GAIL) with PPO as the policy optimizer, and leveraging demonstration data to shape PPO reward functions have shown promising results. A systematic review of these integration patterns can inform future research directions and identify best practices.

Third, the existing surveys employ traditional narrative review methodologies rather than systematic literature review protocols. While narrative surveys provide valuable synthesis and expert interpretation, they may introduce selection bias and limit reproducibility. A systematic approach following established methodologies such as Kitchenham's guidelines [22] and reporting standards such as SEGRESS [23] ensures comprehensive coverage, transparent selection criteria, and reproducible findings.

2.4. Contribution of This Study

This systematic literature review addresses the identified gap by providing focused, methodologically rigorous analysis of PPO applications in autonomous driving. Specifically, this study contributes:

1) Systematic identification and quality assessment of primary studies from 2021–2025 examining PPO for autonomous vehicles;
2) Detailed categorization of PPO implementation approaches, distinguishing standard PPO, modified variants, and baseline comparisons;
3) Comprehensive mapping of imitation learning integration patterns with PPO-based systems;

4) Systematic documentation of evaluation environments, driving tasks, and performance metrics; and
5) Evidence-based identification of research trends, challenges, and future directions specific to PPO in autonomous driving.

By narrowing the scope to PPO while maintaining methodological rigor, this review complements existing broad surveys with the depth necessary for researchers and practitioners seeking to apply or extend PPO-based approaches in autonomous driving systems.

Collectively, these surveys confirm that RL and IL are pivotal in modern AV research and provide excellent, broad overviews of the available methods, applications, and challenges. However, while they all acknowledge PPO as a prominent and effective algorithm, none offer a systematic literature review focused specifically on its application in AV. Given PPO's reputation for stability and sample efficiency in complex control tasks, a dedicated review is necessary to understand precisely how it is being applied (RQ1), its relationship with IL techniques (RQ2), the contexts in which it is evaluated (RQ3), and the metrics used to measure its performance (RQ4). This paper fills this gap by providing a systematic, evidence-based synthesis of the literature on PPO for autonomous driving. Unlike narrative surveys that may selectively review literature, a systematic literature review follows a rigorous, reproducible methodology to minimize bias and provide comprehensive coverage of the available evidence [22].

3. Methods

To investigate the advancement of autonomous vehicles through Proximal Policy Optimization and imitation learning, we conducted a Systematic Literature Review (SLR). This review was designed following established guidelines that distinguish between the *methodology* for conducting the review and the *reporting standards* for documenting it.

For the **review methodology**, we followed the three-stage process proposed by Kitchenham [22], which comprises: (1) planning the review, including defining research questions and the review protocol; (2) conducting the review, including searching, selecting, and extracting data from primary studies; and (3) reporting the review results. This methodology is widely adopted in software engineering research for its systematic and rigorous approach to evidence synthesis.

For the **reporting standards**, we adhered to the SEGRESS (Software Engineering Guidelines for REporting Secondary Studies) framework [23], which provides specific guidance for reporting systematic reviews in software engineering. SEGRESS is an adaptation of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [24] tailored for software engineering contexts. We used the PRISMA flow diagram to visualize our study selection process, as recommended by both frameworks.

Table 3 clarifies the role of each framework in this study.

3.1. Protocol and Registration

A formal protocol was not pre-registered in a public repository prior to conducting this systematic literature review. This decision was made because the review was conducted as part of an academic research project with a predefined scope, and the methodology was established before the search commenced. The complete methodology detailed in the following subsections serves as the full description of the procedures followed for this study, ensuring transparency and reproducibility.

3.2. Eligibility Criteria

To ensure a focused and relevant selection of studies, we defined a specific set of inclusion and exclusion criteria. These criteria were applied to every paper identified during the search process.

3.2.1. Inclusion Criteria. To be included in this review, a study must meet all of the following criteria:

- **Domain:** The study must address autonomous vehicles, self-driving cars, or automated driving systems.
- **Methodology:** The study must discuss reinforcement learning and imitation learning approaches.
- **Algorithm:** The study must utilize or evaluate PPO or its variants as a primary or comparative algorithm.
- **Scope:** The study must focus on single-agent reinforcement learning; multi-agent reinforcement learning (MARL) studies are excluded to maintain a focused scope on single-agent driving policies. MARL introduces additional complexity including communication protocols, coordination mechanisms, and emergent behaviors that warrant separate systematic examination. Furthermore, MARL variants like Multi-Agent Proximal Policy Optimization (MAPPO) represent algorithmic extensions rather than

Table 2. Comparison of Existing Surveys and This Systematic Literature Review

Characteristic	Kiran et al.[a]	Zhu & Zhao[b]	Zhao et al.[c]	This SLR
Methodology	Narrative	Narrative	Narrative	Systematic (Kitchenham + SEGRESS)
Algorithm Scope	All DRL	All DRL + DIL	All DRL	PPO-specific
PPO Coverage	Brief	2 studies	Variants noted	26 studies
IL Integration	Limited	Comprehensive	Moderate	PPO-IL intersection
Quality Assessment	×	×	×	✓
Time Period	2020	2020	2023	2021–2025
Reproducibility	Limited	Limited	Limited	High

[a][19]; [b][20]; [c][21]. DRL = Deep Reinforcement Learning; DIL = Deep Imitation Learning.

Table 3. Distinction Between Frameworks Used in This Study

Framework	Purpose	Application in This Study
Kitchenham [22]	Methodology	Guided how the SLR was planned and conducted (three-stage process)
SEGRESS [23]	Reporting	Guided how the SLR is documented and structured
PRISMA [24]	Visualization	Flow diagram template for study selection process

direct applications of PPO in its standard form.

- **Study Type:** The study must be a peer-reviewed primary study, such as a conference paper or journal article, that presents original empirical results or a novel technique.
- **Publication Date:** The study must have been published between January 1, 2021, and the search execution date of July 7, 2025, to ensure the review focuses on recent advancements.
- **Language:** The study must be written in English.

3.2.2. Exclusion Criteria. A study was excluded if it met any of the following criteria:

- **Domain:** The study does not address autonomous vehicles, self-driving cars, or automated driving systems.
- **Methodology:** The study does not discuss reinforcement learning and imitation learning.
- **Algorithm:** The study does not mention PPO or its variants.
- **Non-Primary Studies:** The publication is a secondary study (e.g., systematic review, survey, or mapping study), an editorial, a keynote, or a tutorial summary.
- **Grey Literature:** The publication is non-peer-reviewed literature, such as university theses, dissertations, blog posts, preprints without peer review, or presentation slides.
- **Full-Text Unavailability:** The full text of the paper could not be accessed through our institutional subscriptions or public archives.
- **Language:** The study is not written in English.
- **Duplicate Reports:** The study reports on the exact same experiment and results as another, more comprehensive paper already included in the review (in such cases, the most complete version was retained).

3.3. Information Sources

To identify relevant primary studies for this review, a comprehensive search was conducted across several major scientific digital libraries known for their extensive coverage of computer science, software engineering, and robotics research. The selected databases were:

- IEEE Xplore Digital Library
- Scopus
- ScienceDirect

These databases were selected based on their comprehensive coverage of peer-reviewed publications in the autonomous driving and machine learning domains, as well as their support for advanced Boolean search queries.

3.4. PICOC Framework

To structure our research scope and derive a systematic search strategy, we applied the PICOC framework, which is widely recommended for formulating research questions in evidence-based software engineering systematic reviews [22]. PICOC provides a structured approach to define the key

elements of a review: Population, Intervention, Comparison, Outcome, and Context. Table 4 presents the PICOC elements defined for this systematic literature review.

Table 4. PICOC Framework for This Systematic Review

Element	Description
Population	Primary studies on autonomous vehicles, self-driving cars, or automated driving systems
Intervention	Application of PPO or its variants within reinforcement learning frameworks
Comparison	Imitation Learning techniques, including BC, GAIL, and AIRL
Outcome	Performance metrics assessing driving policy effectiveness (e.g., collision rate, success rate)
Context	Simulation environments (e.g., CARLA, Highway-Env) and driving tasks (e.g., urban driving, lane changing)

The search string was systematically derived from the PICOC elements to ensure comprehensive coverage of relevant literature. Table 6 illustrates the mapping between PICOC elements and the corresponding search terms.

Table 5. Mapping of PICOC Elements to Search Terms

PICOC Element Inclusion	Search Terms
Population	"Autonomous vehicles" OR "self-driving cars" OR "automated driving"
In query	
Intervention	"Proximal Policy Optimization" OR "PPO"
In query	
Comparison	"Reinforcement Learning" AND "Imitation Learning"
In query	
Outcome	(Performance, effectiveness, metrics)
Screening	
Context	(Simulation, environment, driving task)
Screening	

The Population, Intervention, and Comparison elements were directly incorporated into the Boolean search query using appropriate synonyms and logical operators. The Outcome and Context elements were intentionally excluded from the search string to maximize recall and avoid missing relevant studies that may use varying terminology for performance metrics or evaluation environments. Instead, these elements were operationalized through the eligibility criteria defined in Section 3.2, which required studies to present empirical evaluation results (Outcome) conducted in autonomous driving contexts (Context).

This PICOC-driven approach ensures that our search strategy is both systematic and traceable, directly linking the research questions to the search execution while maintaining sufficient breadth to capture the diverse literature in this interdisciplinary domain.

3.5. Search Strategy

Our search strategy was designed to ensure comprehensive coverage of the relevant literature while maintaining focus on the research scope. The strategy was systematically derived from the PICOC framework defined in Section 3.4 to ensure traceability between our research questions and the search execution.

The primary search was conducted using a predefined search string constructed from the PICOC elements. Population terms ("autonomous vehicles", "self-driving cars", "automated driving") were combined using the OR operator to capture variations in terminology. The Intervention element ("Proximal Policy Optimization", "'PPO") was connected with the AND operator to ensure all results relate to the target algorithm. The Comparison element required both "Reinforcement Learning" and "Imitation Learning" terms to appear, focusing the review on studies addressing the intersection of these paradigms as specified in RQ2.

The Outcome and Context elements from the PICOC framework were intentionally excluded from the search string to maximize recall. Including specific metric names (e.g., "collision rate", "success rate") or environment names (e.g., "CARLA" , "Highway-Env") would have over-constrained the search and potentially excluded relevant studies using different terminology. Instead, these elements were assessed during the screening phase through the eligibility criteria defined in Section 3.2.

The final search string, adapted to the syntax requirements of each database, is as follows:

```
("Autonomous vehicles" OR "self-driving cars" OR "automated driving") AND ("Reinforcement Learning" AND "Imitation Learning") AND ("Proximal Policy Optimization" OR "PPO")
```

Table 6 summarizes the mapping between PICOC elements and their operationalization in the search strategy.

The search was executed on July 7, 2025, across the digital libraries specified in Section 3.3 (IEEE Xplore, Scopus, and ScienceDirect). No date restrictions were applied at the search level; instead, the publication date criterion (2021–2025) was applied during the screening phase to ensure consistent application across all databases.

Table 6. Mapping of PICOC Elements to Search Strategy

Element Applied In	Search Terms
Population	”Autonomous vehicles” OR ”self-driving cars” OR ”automated driving”
Query Intervention	”Proximal Policy Optimization” OR ”PPO”
Query Comparison	”Reinforcement Learning” AND ”Imitation Learning”
Query Outcome	Performance metrics, effectiveness
Screening Context	Environments, driving tasks
Screening	

The use of the conjunctive operator (AND) between ”Reinforcement Learning” and ”Imitation Learning” represents a deliberate methodological trade-off between recall and precision. This design choice prioritizes **precision** (ensuring all retrieved studies are highly relevant to the RL-IL intersection) over **recall** (capturing all PPO studies regardless of IL involvement). Specifically, this approach focuses the review on studies explicitly addressing both paradigms, which is central to RQ2. The trade-off is that studies applying PPO purely within RL frameworks without IL components are excluded. We judged this acceptable because:

1) Our research questions specifically target the RL-IL intersection
2) High precision ensures manageable, focused synthesis
3) Pure RL studies without IL are covered in existing broad surveys [19, 21]

This limitation is further discussed in Section 5.7.

3.6. Selection Process

The study selection was conducted systematically to ensure rigorous and unbiased application of the eligibility criteria defined in Section 3.2.1. The process was managed using the Mendeley reference management tool and performed in two main phases.

3.6.1. Title and Abstract Screening. First, all records identified through the search strategy were imported into the Mendeley reference manager. Automated and manual checks were performed to remove duplicate entries. Following this, the title and abstract of each unique record were screened against the inclusion and exclusion criteria. Each paper was marked as “include” or “exclude.” Papers marked as “uncertain” during initial screening were retained for full-text review. This phase resulted in a list of potentially relevant studies for full-text assessment.

3.6.2. Full-Text Screening. In the second phase, the full-text versions of all papers marked as “include” or “uncertain” from Phase 1 were retrieved. The complete text of each paper was read and assessed against the full set of eligibility criteria to make a final inclusion decision. Reasons for exclusion at this stage were documented to ensure transparency.

3.7. Data Collection Process

To ensure that all relevant information was extracted from the included studies in a consistent manner, we designed a structured data extraction form. The form was developed using a spreadsheet application and was based on the specific requirements of our research questions.

The data extraction was performed by one primary reviewer who systematically read each included paper and populated the data extraction form. To mitigate potential bias from single-reviewer extraction, the following measures were implemented: (1) the extraction form was pilot-tested on three randomly selected studies before full extraction commenced; (2) strict adherence to predefined extraction categories was maintained; and (3) ambiguous cases were flagged and resolved through consultation of the original study text. No contact with the original study authors was required during this process.

3.8. Data Items

The data extraction form was structured to capture the specific data items required to answer our four research questions, along with general bibliographic information for context.

3.8.1. General Study Information. For each included study, the following bibliographic data was extracted:

- Unique study identifier (S1, S2, . . . , Sn)
- Title
- Author(s)
- Publication type (conference paper or journal article)
- Publication venue
- Year of publication

3.8.2. Algorithm and Technique Details. To answer RQ1 and RQ2, the following data points were collected:

- **PPO Application (RQ1):** The specific variant of PPO used (e.g., standard PPO, PPO-Clip, PPO-Penalty, Lagrangian PPO) and a description of how it was applied within the RL framework, including any modifications or enhancements.
- **IL Technique (RQ2):** The name of any imitation learning technique that was discussed, used for comparison, or integrated into the study's methodology (e.g., Behavioral Cloning, DAgger, GAIL, inverse reinforcement learning).

3.8.3. Application and Evaluation Context. To answer RQ3 and RQ4, the following data points were collected:

- **Driving Task and Environment (RQ3):** The autonomous driving task addressed (e.g., lane keeping, lane changing, urban navigation, highway driving, racing) and the simulation environment or platform used for evaluation (e.g., CARLA, SUMO, HighwayEnv, MetaDrive).
- **Performance Metrics (RQ4):** The specific metrics used to evaluate the performance of the driving policy, including their mathematical definitions where provided (e.g., success rate, collision rate, episode reward, time to completion, path deviation).

3.9. Study Risk of Bias Assessment

To evaluate the methodological quality and potential for bias in each included primary study, a risk of bias assessment was performed. This assessment is crucial for interpreting study results and for determining the overall certainty of the evidence in our synthesis.

The assessment was based on a quality checklist composed of four key questions (QA1–QA4), designed to evaluate the rigor and transparency of each study:

- **QA1 – Clarity of Context:** Is the autonomous driving task and the evaluation environment clearly and adequately described?
- **QA2 – Clarity of Methodology:** Is the implementation of the RL/IL algorithm and the experimental setup described in sufficient detail to allow for potential replication?
- **QA3 – Validity of Evaluation:** Are the performance metrics appropriate for the study's stated goals? Is the evaluation conducted against a meaningful baseline or state-of-the-art comparison?
- **QA4 – Transparency of Findings:** Are the results presented clearly and unambiguously? Do the authors explicitly discuss the limitations and threats to the validity of their findings?

Each question was answered on a three-point scale: "Yes" (1 point), "Partly" (0.5 points), or "No" (0 points). The assessment for all included studies was performed by a single reviewer to ensure consistent application of the criteria.

Based on the total score (maximum of 4 points), studies were categorized into three quality tiers:

- **High Quality** (Low Risk of Bias): Score $>$ 3
- **Medium Quality** (Moderate Risk of Bias): Score $>$ 2 to $\leq$ 3
- **Low Quality** (High Risk of Bias): Score $\leq$ 2

The results of this assessment inform the certainty assessment of our synthesized findings, as described in Section 3.13.

3.10. Quality Assessment Framework Development

The quality assessment framework was synthesized from established systematic review guidelines to address key methodological concerns relevant to reinforcement learning research in autonomous driving. While no single existing tool comprehensively addresses the specific needs of assessing machine learning empirical studies, the four assessment questions were designed to align with core quality principles established in systematic review methodology literature.

3.10.1. Derivation of Assessment Questions.
QA1: Clarity of Context. "Is the autonomous driving task and the evaluation environment clearly and adequately described?"

This question addresses the adequacy of contextual description emphasized by Kitchenham and Charters [25] in their systematic review guidelines, which stress that readers must be able to understand the applicability and generalizability of research findings. The question is essential for autonomous driving research where performance is highly context-dependent (e.g., highway vs. urban scenarios, weather conditions, traffic density). This aligns with PRISMA 2020 requirements for reporting study settings and with SEGRESS [23] emphasis on contextual transparency.

QA2: Clarity of Methodology. "Is the implementation of the RL/IL algorithm and the experimental setup described in sufficient detail to allow for potential replication?"

This question addresses reproducibility requirements established in Kitchenham and Charters [25] and reinforced by SEGRESS [23], ensuring independent verification is feasible. Reproducibility is a fundamental concern in machine learning research, as noted in reproducibility checklists developed for venues such as NeurIPS and ICML. The question assesses whether critical implementation details (network architectures, hyperparameters, training procedures) are sufficiently documented for replication.

QA3: Validity of Evaluation. "Are the performance metrics appropriate for the study's stated goals? Is the evaluation conducted against a meaningful baseline or state-of-the-art comparison?"

This question addresses outcome assessment rigor adapted from Cochrane Risk of Bias principles and quality concerns identified in software engineering systematic review research [26]. It ensures that performance claims are properly supported by appropriate metrics and meaningful comparisons. The question recognizes that evaluation validity in autonomous driving research depends both on metric appropriateness (e.g., using safety metrics for safety claims) and on baseline quality (comparisons against relevant alternatives rather than weak strawmen).

QA4: Transparency of Findings. "Are the results presented clearly and unambiguously? Do the authors explicitly discuss the limitations and threats to the validity of their findings?"

This question addresses reporting transparency requirements emphasized by SEGRESS [23] and contemporary reproducibility standards in machine learning research. Limitation discussion is critical for enabling readers to assess the boundary conditions of reported findings and avoid inappropriate generalization. The question also encompasses code and data availability, which are increasingly recognized as essential for reproducible research [23].

These four questions were synthesized to provide a focused assessment appropriate for the autonomous driving reinforcement learning context, while maintaining alignment with established systematic review quality principles. The questions address the most critical dimensions of methodological rigor identified across multiple systematic review frameworks, adapted to the specific characteristics of machine learning empirical research.

3.11. Analysis and Synthesis Methods

To answer the research questions, the data extracted from the included studies were analyzed and synthesized using quantitative methods. The specific approach was chosen based on the nature of each research question.

3.11.1. Synthesis for Research Questions. For RQ1, RQ2, RQ3, and RQ4, which focus on identifying the distribution of techniques, tasks, environments, and metrics, a quantitative synthesis was performed. We used frequency analysis to count the occurrences of specific PPO applications, IL techniques, driving tasks, simulation environments, and performance metrics across the set of included studies.

The results were aggregated and presented visually using tables and bar charts to provide a clear mapping of the research landscape. Descriptive statistics (frequencies and percentages) were calculated to characterize the distribution of each variable of interest.

3.12. Reporting Bias Assessment

We acknowledge the potential for reporting bias (publication bias), where studies with positive or significant results may be more likely to be published than those with negative findings. The presence of such bias could skew the overall conclusions of our synthesis.

As this systematic literature review does not perform a quantitative meta-analysis, formal statistical methods for assessing publication bias (e.g., funnel plots, Egger's test) are not applicable. Instead, our primary method for mitigating this risk was to design and execute a comprehensive search strategy across multiple databases, as detailed in Section 3.5.

By combining searches across multiple major digital libraries, we aimed to identify as much of the relevant literature as possible, regardless of the direction of findings. While no search can be guaranteed to be completely exhaustive, we are confident that our rigorous process has minimized the risk of reporting bias influencing the results of this review.

3.13. Certainty Assessment

To determine the overall confidence in the body of evidence for each synthesized finding, a certainty assessment was performed after data synthesis was complete. This assessment evaluated the extent to which we can be confident that the findings from our synthesis are a correct representation of the available evidence.

For our quantitative findings (RQ1–RQ4), the certainty of findings related to the frequency and distribution of algorithms, techniques, and evaluation practices was assessed based on two main factors:

1) **Comprehensiveness of the Evidence Base:** The total number of studies contributing to each finding.
2) **Overall Risk of Bias:** The distribution of quality assessment scores among the contributing studies, with higher proportions of high-quality studies increasing certainty.

Findings supported by a larger number of high-quality studies were assigned higher certainty, while findings derived from fewer studies or studies with methodological limitations were assigned lower certainty and interpreted with appropriate caution.

4. Results

This section presents the findings of our systematic literature review. The execution of our search strategy across all specified information sources initially yielded a total of 108 records. After a two-phase selection process involving duplicate removal and screening of titles, abstracts, and full texts against our eligibility criteria, a final set of 26 primary studies were included in the synthesis. The following subsections provide a detailed account of this selection process, an overview of the characteristics of the included studies, the results of the risk of bias assessment, and the synthesized findings for each research question.

4.1. Study Selection

The execution of our search strategy across all specified information sources yielded a total of 108 records: 72 from IEEE Xplore, 28 from ScienceDirect, and 8 from Scopus. Table 7 provides a detailed breakdown of the selection process at each phase.

Pre-screening Phase. After importing all records into the Mendeley reference manager, 4 duplicate records were identified and removed. An additional 3 records could not be accessed through our institutional subscriptions. This resulted in 101 unique records available for screening.

Phase 1: Title and Abstract Screening. The titles and abstracts of all 101 records were screened against the eligibility criteria. Records that were clearly outside the scope were excluded at this stage. A total of 52 records were excluded: 17 were not in the autonomous vehicles domain, 24 did not mention imitation learning, 8 were secondary studies (surveys or reviews), and 3 focused on multi-agent reinforcement learning. This phase resulted in 49 potentially relevant studies advancing to full-text assessment.

Phase 2: Full-text Assessment. The full texts of the 49 candidate studies were retrieved and assessed against the complete eligibility criteria. A total of 23 studies were excluded: 6 upon closer examination were not primarily focused on the AV domain, 10 did not substantively discuss imitation learning (only brief mentions in related work), 5 were identified as secondary studies, and 2 were found to focus on multi-agent scenarios.

Table 7. Study Selection Statistics by Phase

Selection Phase	Count
Identification	
Records from IEEE Xplore	72
Records from ScienceDirect	28
Records from Scopus	8
Total records identified	**108**
Pre-screening removal	
Duplicate records removed	4
Records not accessible	3
Records for screening	**101**
Phase 1: Title & Abstract Screening	
Excluded (not AV domain)	17
Excluded (no IL mention)	24
Excluded (secondary study)	8
Excluded (multi-agent RL)	3
Total excluded in Phase 1	**52**
Passed to full-text review	**49**
Phase 2: Full-text Assessment	
Excluded (not AV domain)	6
Excluded (no IL mention)	10
Excluded (secondary study)	5
Excluded (multi-agent RL)	2
Total excluded in Phase 2	**23**
Studies included in synthesis	**26**

This rigorous two-phase selection process resulted in a final set of **26 primary studies** included in the synthesis. The complete selection process is illustrated in the PRISMA flow diagram in Figure 1.

Summary of Exclusions. Across both phases, the total exclusions by reason were: 23 not in AV domain (17+6), 34 no IL mention (24+10), 13 secondary studies (8+5), and 5 multi-agent RL (3+2), totaling 75 excluded records.

4.2. Study Characteristics

This section provides an overview of the general characteristics of the 26 primary studies included in this review. A detailed summary of each individual study, including its publication venue, year, and the specific learning method used, is presented in Appendix A. The following paragraphs summarize the key trends observed across the evidence base.

4.2.1. Publication Type. The majority of the included studies ($n = 21$, 80.8%) were published in peer-reviewed conference proceedings. The remaining 5 studies (19.2%) were published in archival journals, such as IEEE Transactions on Intelligent

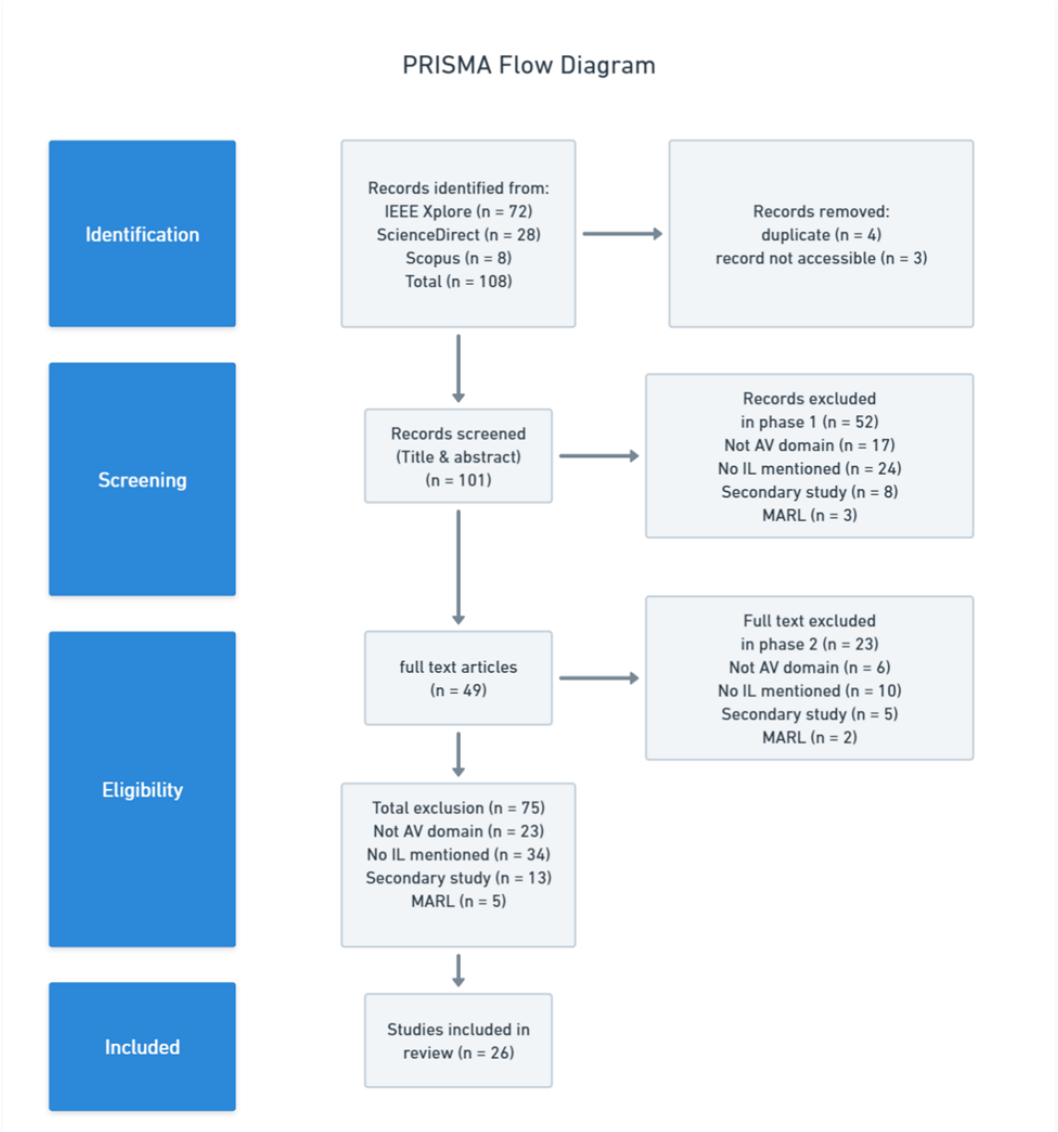

Figure 1. PRISMA flow diagram illustrating the study selection process. The search yielded 108 records, with 26 primary studies included in the final synthesis.

Vehicles and Neurocomputing. Figure 2 shows the distribution of publications by type.

4.2.2. Publication Year. Figure 3 illustrates the distribution of publications by year. The data reveals growing interest in this research area over time. Publications from 2021 account for 6 studies (23.1%), followed by 3 studies in 2022 (11.5%), 5 studies in 2023 (19.2%), 8 studies in 2024 (30.8%), and 4 studies in 2025 (15.4%). Overall, 65.4% of the included studies ($n = 17$) were published between 2023 and 2025, indicating rapidly increasing research activity in recent years.

4.3. Visual Summary of Key Themes

To provide a high-level visual summary of the dominant themes across the included literature, a word cloud was generated from the combined titles and abstracts of all 26 primary studies. As shown in Figure 4, after filtering common academic and topic-specific stopwords, keywords related to evaluation

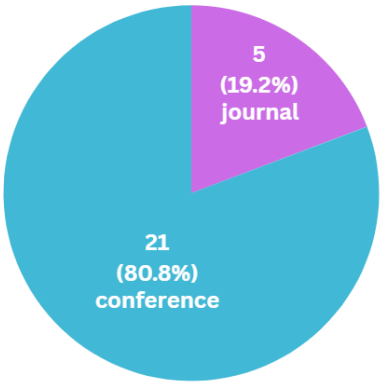

Figure 2. The distribution of publications based on publication type.

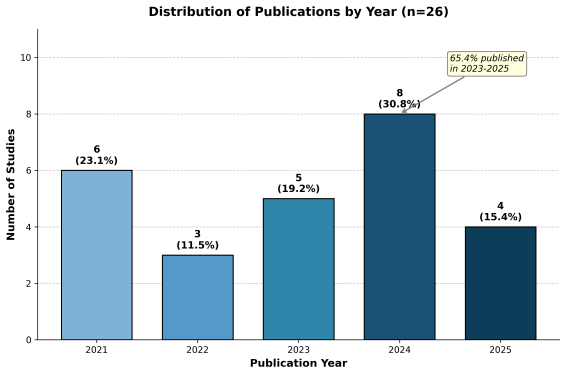

Figure 3. Distribution based on the publication year.

environments, safety, and specific learning architectures emerge prominently.

The importance of terms such as 'environment', 'scenario', 'safety', 'behavior', 'agent', and 'end-to-end' visually reinforces the quantitative findings of this review. These terms indicate a strong community focus on developing and training agents in simulated scenarios where safety and realistic behavior are paramount. Furthermore, the appearance of domain-specific terms such as 'CARLA', 'GAIL', 'adversarial', and 'human' directly reflects the specific tools, algorithms, and concepts that are common in this research area. Section 4.5 provides a detailed quantitative analysis of these themes.

4.4. Risk of Bias in Studies

Each of the 26 included studies was evaluated for risk of bias using the quality assessment checklist detailed in Section 3.9. The assessment was performed by a single reviewer to ensure consistent application of the criteria across all studies. Overall, 12 studies (46.2%) achieved high quality scores, 9 (34.6%) medium quality, and 5 (19.2%) low quality. A detailed breakdown of the scores for each individual study against each quality question (QA1–QA4) is available in Appendix A.

4.5. Results of Syntheses

This section presents the synthesized findings from the 26 included studies, structured to directly answer each of our four research questions.

4.5.1. RQ1: How is PPO being applied within RL frameworks?. Our analysis of the included studies reveals that Proximal Policy Optimization is applied in various forms within autonomous driving research. As illustrated in Figure 5, the distribution of PPO application types is as follows:

The **standard PPO algorithm** is the most frequently used implementation, found in 19 studies (73.1%). These studies apply the original PPO algorithm with its clipped surrogate objective function without significant architectural modifications, though they may customize reward functions and neural network configurations for their specific driving tasks.

A notable portion of the literature ($n = 6$, 23.1%) utilizes **modified PPO variants**, involving algorithmic enhancements tailored to autonomous driving requirements. Examples of modifications identified in the included studies include:

- **Adaptive Clipping PPO (ACPPO/ACPPO+):** Dynamically adjusts the clipping range based on sample importance, allowing larger updates when learning from expert demonstrations (S2, S8)
- **Roach:** A modified PPO incorporating exploration loss and specialized training procedures for end-to-end urban driving (S7)
- **T3PO:** Trajectory Predictive PPO integrating multi-step trajectory prediction (S17)
- **Guided PPO (GPPO):** Incorporates expert demonstrations through Answer Set Programming rules (S26)
- **Leaky PPO:** Addresses issues in original PPO versions for improved autonomous driving performance (S13)

One study (3.8%) mentioned PPO but used a different algorithm as the main contribution, employing **PPO as a baseline** for comparative evaluation (S6).

4.5.2. RQ2: Which IL techniques are discussed or used?. To answer RQ2, we analyzed which Imitation Learning (IL) techniques were discussed or utilized within the 26 included PPO-focused studies.

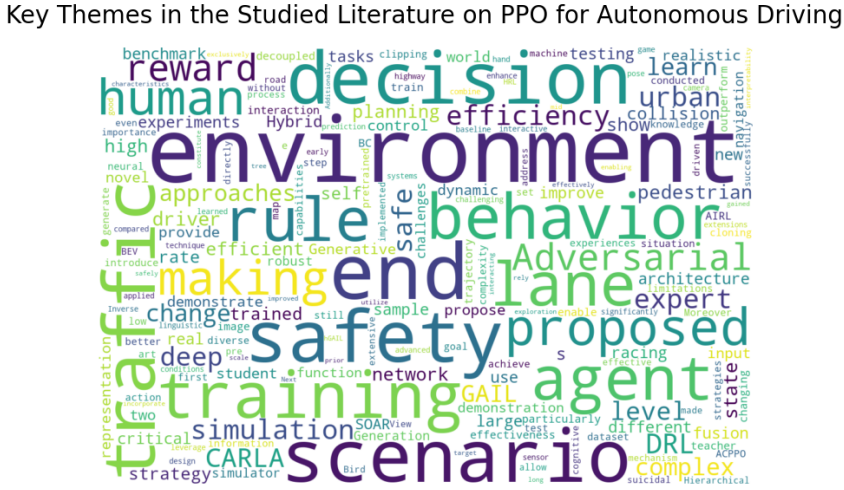

Figure 4. Word cloud generated from the titles and abstracts of the 26 included studies, highlighting the most frequent and significant terms.

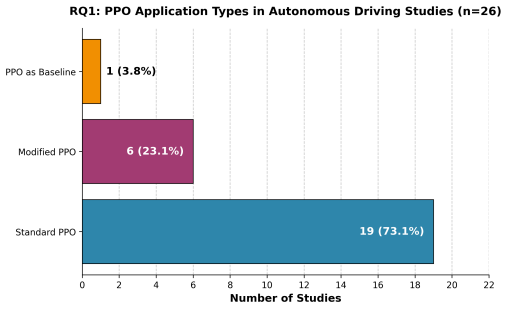

Figure 5. Distribution of PPO application types across the 26 included studies.

As illustrated in Figure 6, the findings reveal diverse approaches to IL integration.

Behavioral Cloning (BC) appeared in 3 studies (11.5%) as the primary IL technique. BC was typically used for policy pre-training, where expert demonstrations initialize the PPO agent before fine-tuning through reinforcement learning. This approach addresses the cold-start problem and accelerates convergence.

Generative Adversarial Imitation Learning (GAIL) was found in 3 studies (11.5%). GAIL-based approaches use PPO as the policy optimizer within the adversarial framework, learning to match the state-action distribution of expert demonstrations.

Adversarial Inverse Reinforcement Learning (AIRL) appeared in 3 studies (11.5%), used for learning reward functions from expert demonstrations that can then guide PPO training.

Multiple IL techniques were combined in 5 studies (19.2%), including combinations such as BC with GAIL, BC with DAgger, and other hybrid approaches.

A substantial number of studies ($n = 6$, 23.1%) employed **other IL techniques** not falling into the above categories. These included:

- Reinforcement Learning from Demonstrations (RLfD)
- Joint Imitation-Reinforcement Learning (JIRL) frameworks
- Teacher-student learning with knowledge transfer
- Hybrid reward shaping using demonstration data
- Custom imitation learning architectures (e.g., LORD, SafeVIPER)

Notably, **6 studies (23.1%)** did not specify or utilize IL techniques despite appearing in our search results. This finding indicates that some studies mention both RL and IL in their text (e.g., in related work sections or as baseline comparisons) without actively integrating IL into their proposed methodology. This is a consequence of our search string design, which required both terms to appear in the paper, as discussed in Section 5.7.

4.5.3. RQ3: What are the common tasks and environments?. Our analysis of the included studies identified a range of autonomous driving tasks and

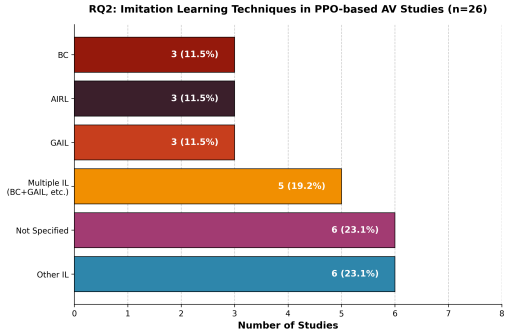

Figure 6. Distribution of Imitation Learning techniques across the 26 included studies.

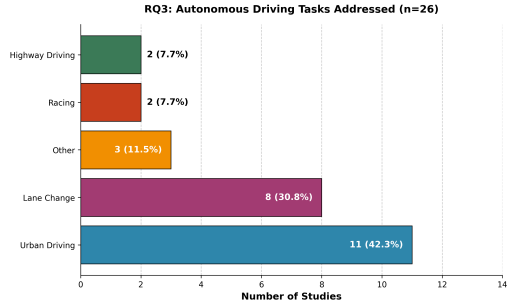

Figure 7. Distribution of autonomous driving tasks addressed in the 26 included studies.

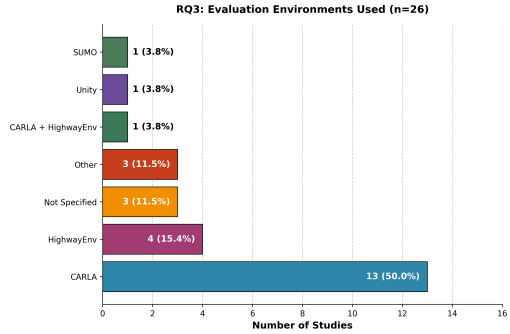

Figure 8. Distribution of evaluation environments used in the 26 included studies.

evaluation environments. As shown in Figures 7 and 8, certain tasks and platforms dominate the research landscape.

Driving Tasks. **Urban driving** was the most frequently targeted task, appearing in 11 studies (42.3%). Urban scenarios typically involve navigation through intersections, traffic light compliance, pedestrian avoidance, and interaction with multiple road users in complex city environments.

Lane changing maneuvers were addressed in 8 studies (30.8%), focusing on lateral decision-making on highways or multi-lane roads. These studies examine safe and efficient lane change execution in the presence of surrounding traffic.

Other tasks identified include:

- **Racing scenarios:** 2 studies (7.7%) – high-speed control in competitive driving environments
- **Highway driving:** 2 studies (7.7%) – longitudinal control and navigation on highways
- **Other tasks:** 3 studies (11.5%) – including lane keeping, continuous control, and obstacle avoidance

Evaluation Environments. The **CARLA simulator** was the dominant evaluation platform, utilized in 13 studies (50.0%). CARLA's open-source nature, realistic urban environments, diverse weather conditions, and extensive sensor suite make it particularly suitable for end-to-end autonomous driving research. One additional study used both CARLA and Highway-Env (3.8%).

Highway-Env was the second most common platform, found in 4 studies (15.4%). This lightweight OpenAI Gym-compatible environment is particularly suited for highway scenarios and behavioral decision-making tasks due to its simplicity and fast simulation speed.

Other environments identified include:

- **SUMO:** 1 study (3.8%) – traffic simulation for lane-changing behavior
- **Unity-based simulators:** 1 study (3.8%) – custom racing environments
- **Other/Custom:** 3 studies (11.5%) – including TORCS, MetaDrive, and custom implementations
- **Not specified:** 3 studies (11.5%) – simulation environment not clearly stated

4.5.4. RQ4: What performance metrics are used?. Our analysis identified 14 unique metrics used across the included studies to evaluate the performance of PPO-based driving policies. This variety indicates a notable lack of standardized evaluation protocols within the research community. A complete list of all identified metrics and their frequencies is provided in Table 8.

As shown in Table 8, metrics related to **safety and task completion** were the most prevalent. **Collision rate** was the most commonly reported metric, appearing in 15 studies (57.7%), reflecting the paramount importance of safety in autonomous driving evaluation. This was followed by **average reward** (10 studies, 38.5%), which serves as a general indicator of policy quality, and **success rate** (8

Table 8. Frequency of Performance Metrics Used in Included Studies ($n = 26$)

Performance Metric	Count	Percentage
Collision Rate	15	57.7%
Average Reward	10	38.5%
Success Rate	8	30.8%
Navigation Distance	5	19.2%
Driving Score	4	15.4%
Center Deviation	3	11.5%
Lane Keeping Accuracy	3	11.5%
Driving Speed	2	7.7%
RMSE	2	7.7%
Reaction Time	2	7.7%
Traffic Light Violation	2	7.7%
Route Completion	1	3.8%
Inter-vehicle Distance	1	3.8%
Lane Violation	1	3.8%

studies, 30.8%), measuring task completion reliability.

Other frequently used metrics included navigation distance (5 studies) for assessing exploration and route completion, and driving score (4 studies), typically a composite metric combining multiple evaluation criteria. Metrics related to driving quality, such as center deviation and lane keeping accuracy, were each found in 3 studies.

The wide range of metrics used presents a significant challenge for the research community, as direct comparison of policy performance across different studies becomes difficult without standardized benchmarks. This issue is explored further in Section 5.

Metric Definitions. To facilitate understanding and support future standardization efforts, we provide formal definitions for the most commonly used metrics:

Collision Rate measures the proportion of episodes in which the autonomous vehicle collides with any object (vehicles, pedestrians, or infrastructure):

$$\text{Collision Rate} = \frac{N_{\text{collisions}}}{N_{\text{episodes}}} \tag{3}$$

where $N_{\text{collisions}}$ is the number of episodes with at least one collision, and N_{episodes} is the total number of test episodes.

Average Reward represents the mean cumulative reward obtained by the agent during evaluation:

$$\text{Average Reward} = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{4}$$

where R_i is the cumulative reward in episode i, and N is the total number of episodes.

Success Rate quantifies how often the vehicle completes a task successfully without collisions or rule violations:

$$\text{Success Rate} = \frac{N_{\text{success}}}{N_{\text{episodes}}} \tag{5}$$

Center Deviation measures the average lateral distance from the lane center:

$$\text{Center Deviation} = \frac{1}{T}\sum_{t=1}^{T} |x_t - x_{\text{center}}| \tag{6}$$

where x_t is the lateral position at time step t, and x_{center} is the lane center position.

Lane Keeping Accuracy evaluates the proportion of time the vehicle remains within lane boundaries:

$$\text{Lane Keeping Accuracy} = \frac{T_{\text{in\_lane}}}{T_{\text{total}}} \tag{7}$$

where $T_{\text{in\_lane}}$ is the time spent within lane boundaries, and T_{total} is the total driving time.

Driving Score is typically a weighted composite metric incorporating multiple evaluation criteria:

$$\text{Driving Score} = \sum_{j=1}^{m} w_j \cdot s_j \tag{8}$$

where w_j is the weight and s_j is the score for the j-th criterion (e.g., safety, efficiency, comfort).

5. Discussion

This section interprets the synthesized findings from our systematic review, situates them within the broader context of autonomous driving research, and discusses their implications for researchers and practitioners. We also acknowledge the limitations of this study and propose directions for future research.

5.1. Summary of Evidence

Our systematic literature review identified 26 primary studies published between 2021 and 2025 that apply Proximal Policy Optimization to autonomous driving tasks while engaging with imitation learning concepts. The evidence reveals several key findings.

Regarding PPO application (RQ1), the standard PPO algorithm dominates the field, with 73.1% of studies (19/26) applying the original implementation without significant architectural modifications. However, a growing subset of research (23.1%, 6/26) develops modified PPO variants specifically tailored to autonomous driving requirements, including adaptive clipping mechanisms (ACPPO), trajectory prediction integration (T3PO), and expert-guided approaches (GPPO). This suggests that while

PPO's core algorithm provides a stable foundation, researchers increasingly recognize the need for domain-specific enhancements.

For imitation learning integration (RQ2), our findings reveal a diverse landscape of IL techniques combined with PPO-based frameworks. Behavioral Cloning, GAIL, and AIRL each appeared in 11.5% of studies, while 19.2% employed multiple IL techniques in combination. Notably, 23.1% of studies did not actively integrate IL despite appearing in our search results, a consequence of our search strategy that required both terms but captured studies mentioning IL only in related work or as baselines. This finding itself highlights an important methodological consideration for future systematic reviews in this domain.

Concerning evaluation environments and tasks (RQ3), the CARLA simulator emerged as the dominant platform (50.0%), with urban driving scenarios (42.3%) and lane changing tasks (30.8%) being the most frequently addressed. This concentration reflects both the maturity of CARLA as a research platform and the community's focus on complex, realistic driving scenarios that require sophisticated decision-making.

Regarding performance metrics (RQ4), our analysis identified 14 unique metrics across the included studies, with collision rate (57.7%), average reward (38.5%), and success rate (30.8%) being most prevalent. This diversity presents both opportunities and challenges, as discussed in subsequent subsections.

5.2. Interpretation of Findings

5.2.1. The Prevalence of Standard PPO. The dominance of standard PPO in autonomous driving research (73.1%) can be attributed to several factors. First, PPO's clipped surrogate objective provides inherent stability guarantees that are particularly valuable in safety-critical applications where unpredictable policy updates could lead to catastrophic failures. Second, the algorithm's relatively low sample complexity compared to other policy gradient methods aligns well with the computational constraints of driving simulation, where each episode requires substantial resources. Third, PPO's implementation simplicity, with robust open-source libraries such as Stable Baselines3 and RLlib, lowers the barrier to entry for researchers exploring autonomous driving applications.

However, the emergence of modified PPO variants (23.1%) signals growing awareness that vanilla PPO may be insufficient for the unique challenges of autonomous driving. Modifications such as Adaptive Clipping PPO (ACPPO) address the tension between exploration and exploitation when learning from expert demonstrations, while trajectory-predictive variants (T3PO) incorporate multi-step planning crucial for anticipatory driving behavior. These trends suggest an evolution toward more specialized algorithms that retain PPO's stability while addressing domain-specific requirements.

5.2.2. The Role of Imitation Learning. Our findings reveal that imitation learning serves multiple complementary roles in PPO-based autonomous driving systems. When used for pre-training (as with BC in 11.5% of studies), IL addresses the cold-start problem by providing reasonable initial policies that can be refined through reinforcement learning. When integrated into the training loop (as with GAIL and AIRL, each at 11.5%), IL shapes the learning objective to match expert behavior distributions, potentially improving sample efficiency and policy quality.

The diversity of IL techniques observed including RLfD, teacher-student learning, and hybrid reward shaping indicates that researchers are actively exploring the optimal way to combine demonstration data with PPO's policy optimization. This experimentation is valuable, but the lack of systematic comparisons between different integration approaches makes it difficult to identify best practices. Future research should prioritize controlled ablation studies that isolate the contributions of different IL components within PPO frameworks.

5.2.3. Simulation Dominance and the Sim-to-Real Gap. The overwhelming reliance on simulation, with CARLA accounting for 50.0% of evaluation environments, reflects both the maturity of available simulation tools and the practical impossibility of training RL agents directly on real vehicles. CARLA's dominance is well-justified: it offers photorealistic rendering, realistic physics, diverse weather conditions, and an extensive sensor suite that closely approximates real-world perception challenges.

However, this concentration raises important concerns. First, the sim-to-real gap (the discrepancy between simulation and real-world dynamics) remains a fundamental challenge for deploying RL-trained policies. Only a small number of included studies addressed this gap explicitly through domain randomization or transfer learning techniques. Second, CARLA's specific characteristics (its physics engine, rendering pipeline, and scenario definitions) may inadvertently bias the research community toward solutions that perform well in CARLA but may not generalize to other simulators or real-world

conditions. Third, the relative scarcity of studies using alternatives such as Highway-Env (15.4%), SUMO, or custom environments limits the diversity of evaluation scenarios.

5.2.4. The Metric Fragmentation Problem. Perhaps the most significant finding from our review is the fragmentation of performance metrics, with 14 unique metrics identified across only 26 studies. While collision rate (57.7%) and success rate (30.8%) provide intuitive safety and task completion measures, the proliferation of study-specific metrics, including various formulations of driving score, lane keeping accuracy, and navigation distance, severely limits the ability to compare results across different works.

This fragmentation is not merely an inconvenience; it actively impedes scientific progress. Without standardized evaluation protocols, it becomes impossible to establish whether a novel algorithm genuinely advances the state of the art or simply performs well on a narrowly defined metric. The absence of common benchmarks also makes it difficult to identify systematic weaknesses in existing approaches that might guide future research directions.

5.3. Temporal Evolution of PPO Applications (2021–2025)

Analysis of the 26 included studies reveals clear temporal trends in how PPO is being applied to autonomous driving, with progressive sophistication in algorithmic modifications and integration strategies over the five-year period.

5.3.1. Early Period (2021–2022): Establishing Foundations. The early period comprises 9 studies (34.6%) that primarily established PPO's viability for autonomous driving while beginning to explore modifications and hybrid approaches. Among these, 4 studies (44.4%) used standard PPO implementations, while 5 studies (55.6%) introduced early modifications.

Representative examples from this foundational period include:

S7 [27] introduced Roach, an RL expert using modified PPO with exploration loss for end-to-end urban driving in CARLA. This study demonstrated that RL-trained policies could achieve state-of-the-art performance on the CARLA Leaderboard and NoCrash-dense benchmark, establishing an important baseline for subsequent imitation learning research. The addition of exploration loss helped address the challenge of diverse scenario coverage.

S20 [28] proposed SafeVIPER, combining PPO with verified rule-based controllers for urban navigation. This hybrid approach recognized early that pure RL policies may be insufficient for safety-critical deployment, incorporating formal safety guarantees alongside learned policies. The framework achieved balanced performance between exploration and safety constraints in complex urban scenarios.

S22 [29] explored Generative Adversarial Imitation Learning (GAIL) with PPO as the policy optimizer for end-to-end urban driving. This study demonstrated the benefits of adversarial IL techniques over simple behavioral cloning, reducing collision rates in CARLA urban environments. The use of PPO within GAIL established a pattern of treating PPO as a reliable policy optimization backend for IL frameworks.

S18 [30] applied meta-reinforcement learning with PPO for lane-changing maneuvers in Highway-Env, achieving improved generalization to new environments. This early work recognized that standard PPO required enhancement for transfer learning across diverse driving conditions.

5.3.2. Middle Period (2023): Algorithmic Innovations. The 2023 period (5 studies, 19.2%) shows a marked shift toward algorithmic modifications, with 4 of 5 studies (80%) employing modified PPO variants or novel integration strategies—a substantial increase from the 55.6% in 2021–2022.

Key innovations during this period include:

S16 [31] applied Adversarial Inverse Reinforcement Learning (AIRL) with PPO for traffic simulation in Highway-Env. Rather than hand-crafting reward functions, AIRL learned reward models from naturalistic human driving data, producing more human-like lane-changing behaviors. This represented a shift from reward engineering to reward learning approaches.

S17 [32] introduced T3PO (Trajectory Predictive PPO), embedding trajectory prediction directly into the RL framework. This modification addressed PPO's challenge of planning in dynamic environments by incorporating multi-step prediction, enabling more anticipatory driving policies.

S26 [33] developed Guided PPO (GPPO) using Answer Set Programming to inject expert knowledge from answer sets into PPO training. This knowledge-guided approach combined symbolic reasoning with neural policy learning, targeting complex urban intersection scenarios where standard PPO struggled.

S25 [34] proposed an efficient lane-changing algorithm combining IL initialization with PPO fine-tuning, reducing training time while achieving low collision rates in highway scenarios. This two-stage

approach acknowledged both PPO's sample inefficiency and its strength for policy refinement.

5.3.3. Recent Period (2024–2025): Integration Maturity. The most recent period (12 studies, 46.2%) demonstrates increasing maturity through sophisticated algorithmic modifications and refined integration with imitation learning. Modified variants now represent 6 of 12 studies (50%), with growing emphasis on adaptive mechanisms and domain-specific enhancements.

Notable recent contributions include:

S2 [35] introduced SOAR-ACPPO, integrating a cognitive architecture (SOAR) with Adaptive Clipping PPO for lane changing in CARLA. ACPPO dynamically adjusts the clipping parameter based on sample importance, specifically whether samples originate from expert policies or the agent's own exploration. This allows larger updates when learning from high-quality expert demonstrations while maintaining stability. The framework demonstrated improved learning efficiency and safety compared to both standalone SOAR rules and traditional PPO.

S4 [36] developed an adversarial scenario generation framework using naturalistic human driving priors from NGSIM and INTERACTION datasets. The approach uses GAIL to learn realistic human behavior models, then employs PPO to generate safety-critical scenarios that balance adversariality (challenging the AV) with naturalness (avoiding unrealistic collisions). This addressed the problem of generating meaningful test scenarios for AV validation.

S8 [37] proposed the S2CD (Simple-to-Complex Collaborative Decision) framework with ACPPO+ algorithm. A teacher model is first trained rapidly in simplified Highway-Env, then guides a student model learning in high-fidelity CARLA through intervention based on Q-value comparisons. ACPPO+ incorporates samples from both teacher and student policies with adaptive clipping and KL divergence constraints to accelerate knowledge transfer. This curriculum learning approach significantly improved both training efficiency and safety.

S13 [38] introduced GP3Net (Graph-based Prediction and Planning Policy Network), integrating multi-step trajectory prediction directly into PPO's policy network for dense traffic scenarios. This architectural modification enabled anticipatory planning in dynamic environments.

S3 [39] proposed hierarchical GAIL (hGAIL) using a GAN to generate Bird's-Eye View representations from raw camera images, with PPO as the policy optimizer. This achieved 98% success rate in urban intersection navigation across different cities, demonstrating strong generalization.

S14 [40] introduced LORD (Large Models based Opposite Reward Design), leveraging large pre-trained models as zero-shot reward mechanisms for PPO-based driving. This represented an emerging trend of incorporating foundation models into RL reward design.

5.3.4. Observed Trends Across Periods. Three clear trends emerge from this temporal analysis:

1. Progressive shift from standard to modified PPO. Standard PPO usage declined from 44.4% (2021–2022) to 20% (2023) to 41.7% (2024–2025), while modified variants increased correspondingly. This suggests growing recognition that vanilla PPO requires domain-specific enhancements for challenging autonomous driving scenarios.

2. Evolution of IL integration sophistication. IL integration evolved from simple behavioral cloning pre-training (2021–2022) to sophisticated adversarial methods (GAIL, AIRL) and reward learning (2023), and finally to adaptive multi-modal integration with cognitive architectures and curriculum learning (2024–2025). The nature of integration shifted from sequential (pre-train then fine-tune) to concurrent and adaptive.

3. Increasing focus on practical deployment concerns. Early studies emphasized algorithmic feasibility and basic performance. Middle-period work introduced efficiency improvements (training time reduction). Recent studies increasingly address real-world deployment requirements: safety-critical scenario generation (S4), interpretability via cognitive architectures (S2), curriculum learning for safe training (S8), and sim-to-real considerations through domain randomization.

4. Broadening of application scope. While highway lane-changing dominated early work (S1, S18, S22), recent studies tackle more complex scenarios: urban intersections (S26, S3), multi-agent roundabouts (S5), dynamic route planning (S3), and safety-critical event generation (S4). This progression reflects both algorithmic maturity and community confidence in PPO-based approaches.

5.4. Comparative Performance Patterns

While direct quantitative comparison across studies is limited by environmental and methodological heterogeneity, systematic analysis of reported performance outcomes reveals consistent patterns regarding the relative effectiveness of different approaches.

5.4.1. PPO versus Pure Imitation Learning. Multiple studies explicitly compared PPO-based ap-

proaches against pure imitation learning baselines, with consistent findings:

Studies S7, S8, S15, S22, and S25 all reported that PPO outperformed pure behavioral cloning in their respective scenarios. The consistent pattern across diverse tasks (urban driving, lane changing, racing) suggests PPO's exploration capability provides meaningful performance gains over policies limited to mimicking expert demonstrations. Behavioral cloning typically achieved adequate performance in nominal scenarios but struggled with recovery from near-failures or adaptation to distribution shifts—situations where PPO's reinforcement signal enabled corrective learning.

However, this advantage comes with a cost: S1, S8, S18, and S25 all noted PPO's sample inefficiency, with training requiring substantially more environment interactions than IL approaches. This motivated the widespread adoption of hybrid methods in later years.

5.4.2. Hybrid Approaches Outperform Single Paradigms. A particularly consistent finding across studies is the superiority of hybrid RL-IL approaches over either paradigm alone:

S2 (SOAR-ACPPO) demonstrated that combining cognitive architecture guidance with adaptive PPO exceeded the performance of both standalone rule-based SOAR and traditional PPO. The cognitive architecture provided sample-efficient initialization and safety constraints, while PPO enabled learning beyond hand-coded rules.

S7 (Roach + IL) showed that an imitation learning agent trained on Roach's (RL-expert) demonstrations achieved state-of-the-art CARLA Leaderboard performance, surpassing both the RL expert alone and pure IL approaches. This two-stage approach (RL coach $\rightarrow$ IL student) combined PPO's exploration with IL's stability.

S8 (S2CD framework) reported that teacher-student frameworks with guided exploration outperformed both standalone PPO and pure imitation learning across safety and efficiency metrics. The guided exploration accelerated learning while maintaining safety bounds.

S22 (GAIL + PPO) found that adversarial imitation learning (using PPO as the policy optimizer) achieved lower collision rates than behavioral cloning in urban scenarios. The adversarial training addressed BC's distribution mismatch problem.

S25 (IL initialization + PPO) demonstrated that initializing PPO with IL pre-training substantially reduced training time compared to PPO from scratch, while achieving better final performance than pure IL. This sequential integration captured benefits of both approaches.

This pattern strongly suggests that the optimal strategy lies in thoughtful integration rather than committing to a single paradigm. The specific integration mechanism varies—cognitive guidance (S2), staged training (S7, S25), adversarial learning (S22), or curriculum learning (S8)—but the principle of complementarity consistently yields superior results.

5.4.3. Modified PPO Variants Show Domain-Specific Improvements. Studies introducing domain-specific PPO modifications reported substantial improvements over standard implementations in their target scenarios:

S2 (ACPPO) adaptive clipping based on sample importance demonstrated improved learning efficiency and convergence compared to fixed-clipping standard PPO, particularly when learning from mixed expert-agent experience.

S8 (ACPPO+) further refined adaptive clipping with KL divergence constraints for knowledge transfer, achieving better performance than both standard PPO and safe RL variants (CPO, TRPO) in highway scenarios.

S13 (GP3Net) integration of multi-step trajectory prediction into PPO's policy network showed significant improvements in dense traffic scenarios compared to standard PPO, which struggles with anticipatory planning.

S26 (GPPO) knowledge-guided PPO using symbolic reasoning demonstrated superior performance in complex urban intersections compared to standard PPO, which often failed to navigate multi-agent conflict zones.

These modifications address specific limitations of standard PPO (sample importance weighting, multi-step planning, symbolic knowledge integration) rather than proposing general improvements. This trend suggests that PPO's strength lies partly in its modifiability, the algorithm provides a stable foundation that can be enhanced with domain-specific mechanisms.

5.4.4. Limitations of Cross-Study Comparison. We emphasize several critical limitations that constrain quantitative comparison:

Environmental diversity. CARLA versions ranged from 0.9.10 to 0.9.14 across studies, with variations in physics models, sensor simulations, and scenario configurations. Highway-Env, SUMO, and custom environments introduce further variability.

Scenario inconsistency. Even within CARLA, studies used different maps, weather conditions, traffic densities, and task definitions. "Lane changing" in sparse highway traffic (S18) versus dense urban intersections (S26) represents fundamentally different challenges.

Metric heterogeneity. Success rate definitions varied (collision-free completion vs. collision-free + goal-reaching + comfort constraints). Some studies reported single-run results while others provided multi-seed statistical analyses.

Baseline selection. Comparison fairness depends on baseline implementation quality. When studies reported PPO outperforming BC, it remains unclear whether BC was optimally configured (e.g., dataset size, network architecture, data augmentation).

These comparisons should therefore be interpreted as indicative patterns, consistent across multiple independent studies, rather than definitive rankings. The field would benefit from standardized benchmarks (e.g., CARLA Leaderboard) applied uniformly across studies.

5.5. Comparison with Previous Surveys

Our findings both align with and extend those of previous surveys on deep reinforcement learning for autonomous driving.

Kiran et al. [19] provided a comprehensive survey of DRL applications in autonomous driving but did not focus specifically on PPO or its variants. Their observation that DRL research remains predominantly simulation-based is strongly confirmed by our findings, with 100% of included studies evaluating policies in simulated environments. However, our focused analysis reveals the specific extent of CARLA's dominance (50.0%) and the diversity of PPO modifications being developed, details not captured in broader surveys.

Zhu and Zhao [20] introduced a five-mode taxonomy for DRL and IL integration but identified PPO in only two of their reviewed studies. Our systematic review, covering more recent literature (2021–2025) with a targeted search strategy, identified 26 studies that apply PPO in conjunction with IL concepts. This dramatic increase demonstrates PPO's rapid adoption in autonomous driving research and validates the timeliness of our focused analysis.

Zhao et al. [21] discussed PPO variants including PCPO and MAPPO in their recent survey but did not provide systematic quantification of how different PPO variants are applied or how they integrate with IL techniques. Our analysis fills this gap by providing concrete statistics on PPO application patterns (73.1% standard, 23.1% modified, 3.8% baseline) and IL integration approaches.

Unlike these previous surveys, which employed narrative review approaches, our systematic methodology following Kitchenham guidelines [22] with explicit search strategies, eligibility criteria, and quality assessment enables reproducible findings and quantifiable evidence synthesis. This methodological rigor distinguishes our work and provides a foundation for future meta-analyses as the field matures.

5.6. Quality of Evidence

The methodological quality of the included studies, as assessed through our risk of bias evaluation, was generally acceptable but revealed important patterns. While 46.2% of studies achieved high quality ratings (low risk of bias), nearly one-fifth (19.2%) exhibited significant methodological weaknesses.

The most common deficiencies were related to transparency of findings (QA4), with many studies failing to adequately discuss limitations or threats to validity. This pattern is concerning because it may lead to overconfident claims about algorithm performance that do not account for potential confounds. Additionally, some studies lacked sufficient methodological detail (QA2) to enable replication, with absent hyperparameter specifications, unclear reward function definitions, or unavailable source code.

These quality concerns should temper the interpretation of our synthesized findings. The evidence base, while growing, would benefit from more rigorous reporting standards. We recommend that future studies in this domain adopt standardized reporting guidelines and make training code, hyperparameters, and evaluation scripts publicly available.

5.7. Limitations

This systematic review has several limitations that should be considered when interpreting the findings.

5.7.1. Single Reviewer. A primary methodological limitation is the use of a single reviewer for the data extraction stage, as acknowledged in Section 3.7. While the extraction was conducted systematically using predefined data items and a pilot-tested extraction form, the absence of independent verification introduces potential for subjective bias or unintentional errors. Readers should consider this limitation when interpreting the synthesized findings, particularly for categorization decisions that required judgment (e.g., classifying a PPO implementation as "standard" versus "modified").

5.7.2. Search Strategy. Our search strategy, which required the Boolean AND operator between reinforcement learning and imitation learning terms, was a deliberate methodological choice to focus on studies addressing the intersection of these paradigms.

However, this design may have excluded relevant studies that apply PPO to autonomous driving without engaging with imitation learning concepts. The finding that 23.1% of included studies did not actively integrate IL despite mentioning both terms, illustrates the imprecision of text-based search strategies and suggests that our review captures only a subset of the broader PPO-for-autonomous-driving literature.

Future reviews might complement our approach with alternative search strategies: either broadening the search to capture all PPO-autonomous driving studies (accepting lower precision) or employing snowball sampling from the included studies to identify related work that our search may have missed.

Exclusion of Multi-Agent Scenarios. Our deliberate exclusion of MARL studies limits the applicability of our findings to cooperative or competitive multi-vehicle scenarios. As autonomous vehicles increasingly operate in mixed-autonomy traffic environments where vehicle-to-vehicle coordination becomes critical, future systematic reviews should specifically examine multi-agent PPO variants (like MAPPO) and their application to cooperative driving tasks. This represents an important but distinct research direction requiring dedicated analysis.

5.7.3. Temporal Coverage. Our search, conducted in July 2025, captures literature published through that date. Given the rapid pace of advancement in both deep reinforcement learning and autonomous driving, findings from this review may become dated as new algorithms, simulation platforms, and evaluation methodologies emerge. The temporal distribution of included studies (65.4% from 2023–2025) indicates an accelerating research pace, suggesting that periodic updates to this review will be valuable.

5.7.4. Language and Database Limitations. Our review was limited to English-language publications indexed in the specified databases (IEEE Xplore, Scopus, and ScienceDirect). This may have excluded relevant research published in other languages or in venues not indexed by these databases, particularly work from non-English speaking research communities that are active in autonomous vehicle development.

5.7.5. Technology Maturity and Deployment Readiness.. All 26 reviewed studies (100%) conduct evaluation exclusively in simulation environments (CARLA, Highway-Env, SUMO, Unity, F1TENTH gym, or custom simulators), with zero studies demonstrating real-world deployment or even sim-to-real transfer validation. This indicates the technology remains at an early research stage, with significant gaps separating current simulation performance from production-ready deployment. Key challenges remain largely unaddressed in the literature: (1) the sim-to-real transfer gap, as simulators employ simplified physics and sensor models that may not reflect real-world complexity, (2) computational scalability, as the subset of studies reporting training requirements (n=8, 30.8%) indicated 100–1000+ GPU hours per policy, raising questions about deployment on embedded automotive hardware, (3) multi-scenario generalization, as no study demonstrated a single policy generalizing across substantially different driving contexts (urban, highway, parking), and (4) safety certification pathways, with only 3 studies (11.5%) incorporating formal safety constraints or verified controllers. While simulation results are promising, substantial additional work is required to bridge the gap from research prototypes to production autonomous driving systems. Future research should prioritize real-world validation studies, standardized sim-to-real transfer protocols, and hybrid architectures combining learned policies with verified safety layers.

5.7.6. Dataset and Benchmark Standardization. Our systematic review did not extract detailed dataset-level information (e.g., specific datasets used, demonstration sources, data collection methods) as this was outside our defined research questions. However, this represents an important gap: without systematic analysis of whether studies use comparable datasets and standardized benchmarks, it is difficult to determine whether performance differences across studies reflect algorithmic improvements or merely variations in evaluation difficulty. Future systematic reviews should explicitly extract dataset characteristics and benchmark usage to enable more rigorous cross-study comparison.

5.8. Implications

The findings of this systematic literature review have several important implications for researchers, practitioners, and the broader autonomous driving community.

5.8.1. For Researchers.

Standardization of Evaluation Protocols. The fragmentation of performance metrics identified in this review presents a critical challenge. We strongly recommend that the research community work toward establishing standardized benchmark suites for PPO-based autonomous driving research. Such benchmarks should include:

- A core set of mandatory metrics covering safety (collision rate), task completion (success rate), and efficiency (completion time or distance)
- Standardized driving scenarios with defined difficulty levels, including urban navigation, lane changing, and emergency maneuvers
- Clear reporting requirements for environmental conditions, sensor configurations, and training details

The CARLA Leaderboard provides a foundation for such standardization, but adoption remains incomplete. Community-driven efforts to establish and enforce common benchmarks would substantially improve the ability to compare and build upon existing research.

Addressing the Sim-to-Real Gap. The exclusive reliance on simulation in the included studies underscores the pressing need for research focused on policy transfer to real-world environments. Specific directions include:

- Systematic investigation of domain randomization techniques and their effectiveness for autonomous driving tasks
- Development of improved physics models that better capture real-world vehicle dynamics
- Hybrid approaches that combine simulation training with limited real-world fine-tuning
- Benchmarking of sim-to-real transfer methods across multiple simulators and vehicle platforms

Systematic Comparison of IL Integration Approaches. The diversity of imitation learning techniques identified in this review presents both an opportunity and a gap. Future research should prioritize controlled experiments that systematically compare different IL integration strategies (pre-training, reward shaping, adversarial learning) within consistent PPO frameworks and evaluation settings.

5.8.2. For Practitioners.

Algorithm Selection. For practitioners developing autonomous driving systems, our findings suggest that standard PPO provides a robust baseline with well-established stability properties. However, the emergence of domain-specific modifications indicates that performance gains may be available through algorithmic enhancements tailored to specific driving requirements. Practitioners should consider:

- Starting with standard PPO implementations using established libraries
- Incorporating imitation learning pre-training when expert demonstration data is available
- Evaluating modified PPO variants (e.g., ACPPO, GPPO) for applications where standard PPO shows limitations

Simulation Platform Selection. While CARLA's dominance reflects its capabilities, practitioners should consider environment selection based on specific requirements. Highway-Env offers faster iteration cycles for highway-focused applications, while custom Unity-based environments may provide greater flexibility for specialized scenarios. The key recommendation is to validate policies across multiple simulation environments before considering real-world deployment.

Safety Considerations. The prevalence of safety-related metrics (collision rate in 57.7% of studies) reflects appropriate prioritization of safety in autonomous driving research. However, practitioners should note that simulation-based safety evaluation may not capture all real-world risks. We recommend complementing RL-trained policies with traditional safety mechanisms (rule-based overrides, motion planning constraints) as a defense-in-depth approach.

5.8.3. For the Research Community.

Reproducibility and Open Science. The quality assessment revealed that insufficient methodological detail hampers reproducibility in this domain. We call on the research community to embrace open science practices, including:

- Publishing complete source code and trained model weights
- Providing detailed hyperparameter specifications and ablation studies
- Sharing standardized evaluation scripts and scenario definitions
- Documenting negative results and failed approaches

Alternative Simulation Platforms. While CARLA's dominance is well-justified, over-reliance on a single platform carries risks. We encourage exploration of alternative environments, including general-purpose game engines such as Unity (with ML-Agents) or Godot (with Godot-RL), which offer advanced physics engines, high-fidelity rendering, and flexibility for creating custom scenarios. Diversifying evaluation platforms would improve the robustness and generalizability of research findings.

5.9. Future Research Directions

Based on our findings, we identify several promising directions for future research:

5.9.1. Safety-Constrained PPO Variants. While standard PPO provides policy update stability, it does not inherently enforce safety constraints. Constrained PPO variants (e.g., CPO, PCPO, PPO-Lagrangian) that explicitly incorporate safety constraints into the optimization objective represent a promising direction. Future research should systematically evaluate these variants in autonomous driving contexts and develop specialized safety constraints for driving applications.

5.9.2. Multi-Agent and Cooperative Driving. The included studies predominantly focused on single-agent scenarios. As autonomous vehicles increasingly share roads with both human drivers and other autonomous systems, research on multi-agent PPO variants (e.g., MAPPO, IPPO) for cooperative and competitive driving scenarios becomes essential. This includes vehicle-to-vehicle coordination, mixed-autonomy traffic, and negotiation in ambiguous situations.

5.9.3. Interpretability and Explainability. End-to-end driving policies trained with PPO often function as black boxes, making it difficult to understand and validate their decision-making. Future research should explore methods for interpreting PPO-trained policies, including attention visualization, policy distillation into interpretable representations, and formal verification of safety properties.

5.9.4. Real-World Deployment. The complete absence of real-world evaluation in the included studies represents the field's most significant gap. While simulation remains essential for initial development, the ultimate goal of autonomous driving research is real-world deployment. Future research should develop systematic methodologies for safe, incremental real-world testing of RL-trained policies, including confidence-based intervention systems and graceful degradation strategies.

5.9.5. Efficiency and Resource Constraints. Many modified PPO variants introduce computational overhead that may be impractical for embedded vehicle systems. Research on efficient PPO implementations, including model compression, quantization, and hardware-aware optimization, would facilitate the practical deployment of RL-based driving policies.

6. Other Information

6.1. Competing Interest

The authors declare that they have no competing interests.

6.2. Availability of Data, Code, And Other Materials

All supplementary materials for this systematic literature review, including the full list of included studies, the complete data extraction spreadsheet, and the analysis scripts used to generate the figures in this paper, are available upon request.

Appendix

This appendix provides a comprehensive list of the 26 primary studies that were included in this systematic literature review after the full selection process. Table 9 summarizes key bibliographic information for each study, including its unique identifier (code) used throughout this paper, the citation, publication year, publication type, and publication venue.

This appendix provides the detailed results of the risk of bias assessment conducted for each of the 26 primary studies included in this systematic literature review. The assessment was performed using the quality assessment checklist described in Section 3.9 of the main paper.

1. Quality Assessment Questions

Each study was evaluated against four quality assessment questions (QA1–QA4), with responses scored as follows: Yes = 1, Partly = 0.5, No = 0. The maximum possible score is 4 points.

- **QA1 – Clarity of Context:** Does the study clearly describe the autonomous driving task, environment, and problem context?
- **QA2 – Clarity of Methodology:** Does the study provide sufficient detail about the algorithm implementation, hyperparameters, and experimental setup to enable replication?
- **QA3 – Validity of Evaluation:** Does the study use appropriate metrics and evaluation procedures to assess the proposed approach?
- **QA4 – Transparency of Findings:** Does the study discuss limitations, threats to validity, and provide access to code or data?

Table 9. Summary of Included Studies (n = 26)

Code	Paper	Year	Type	Publication Venue
S1	[41]	2021	Conference	IEEE International Conference on Unmanned Systems (ICUS)
S2	[35]	2024	Journal	Neurocomputing
S3	[39]	2024	Journal	IEEE Transactions on Intelligent Vehicles
S4	[36]	2024	Journal	IEEE Transactions on Intelligent Vehicles
S5	[42]	2024	Conference	IEEE Intelligent Vehicles Symposium (IV)
S6	[43]	2025	Conference	Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS)
S7	[27]	2021	Conference	IEEE/CVF International Conference on Computer Vision (ICCV)
S8	[37]	2025	Journal	Advanced Engineering Informatics
S9	[44]	2022	Conference	International Conference Radioelektronika
S10	[45]	2024	Conference	IEEE/SICE International Symposium on System Integration (SII)
S11	[46]	2021	Conference	IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
S12	[47]	2022	Journal	IEEE Transactions on Vehicular Technology
S13	[38]	2024	Conference	AAAI Conference on Artificial Intelligence (AAAI)
S14	[40]	2025	Conference	IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)
S15	[48]	2023	Conference	IEEE Intelligent Vehicles Symposium (IV)
S16	[31]	2023	Conference	IEEE International Conference on Intelligent Transportation Systems (ITSC)
S17	[32]	2023	Conference	IEEE International Conference on Digital Twins and Parallel Intelligence (DTPI)
S18	[30]	2021	Conference	IEEE Intelligent Vehicles Symposium (IV)
S19	[49]	2025	Conference	International Conference on Electronics, Information, and Communication (ICEIC)
S20	[28]	2021	Conference	IEEE Intelligent Vehicles Symposium (IV)
S21	[50]	2024	Conference	International Conference on Computational Intelligence and Network Systems (CINS)
S22	[29]	2021	Conference	IEEE Symposium Series on Computational Intelligence (SSCI)
S23	[51]	2022	Conference	IEEE Intelligent Vehicles Symposium (IV)
S24	[52]	2024	Conference	International Conference on Artificial Intelligence and Computer Applications (ICAICA)
S25	[34]	2023	Conference	IEEE Intelligent Vehicles Symposium (IV)
S26	[33]	2023	Conference	IEEE International Conference on Tools with Artificial Intelligence (ICTAI)

2. Quality Score Classification

Based on the total scores, studies were classified into three quality tiers:

- **High Quality** (Low Risk of Bias): Total Score > 3.0
- **Medium Quality** (Moderate Risk of Bias): Total Score > 2.0 and ≤ 3.0
- **Low Quality** (High Risk of Bias): Total Score ≤ 2.0

3. Assessment Results

Table 10 presents the complete risk of bias assessment results for all 26 included studies.

4. Summary Statistics

The distribution of studies across quality tiers is summarized below:

- **High Quality** (Low Risk of Bias): 12 studies (46.2%) – S2, S3, S4, S6, S7, S8, S9, S10, S11, S20, S22, S25
- **Medium Quality** (Moderate Risk of Bias): 9 studies (34.6%) – S1, S12, S13, S14, S15, S16, S18, S23, S26
- **Low Quality** (High Risk of Bias): 5 studies (19.2%) – S5, S17, S19, S21, S24

5. Score Distribution by Criterion

Table 11 summarizes the score distribution for each quality assessment criterion across all 26 studies.

6. Observations

Analysis of the quality assessment results reveals several important patterns:

QA1 – Clarity of Context. The majority of studies performed well on this criterion, with 22 studies (84.6%) receiving full scores. Most studies clearly described their autonomous driving tasks, simulation environments (e.g., CARLA, Highway-Env), and problem contexts. Only one study (S17) received a score of 0 due to unclear task description and unspecified environment.

QA2 – Clarity of Methodology. Performance on this criterion was more variable. While 16 studies (61.5%) received full scores, 3 studies (11.5%) received scores of 0 (S14, S19, S24). Common deficiencies included:

- Missing or incomplete hyperparameter specifications
- Lack of detailed experimental procedures
- Absence of publicly available source code

Table 10. Risk of Bias Assessment Results ($n = 26$). Scoring: Yes = 1, Partly = 0.5, No = 0. Maximum total score = 4.

Code	QA1	QA2	QA3	QA4	Total	Quality Tier
S2	1	1	1	1	4.0	High
S3	1	1	1	0.5	3.5	High
S4	1	1	1	1	4.0	High
S6	1	1	1	0.5	3.5	High
S7	1	1	1	0.5	3.5	High
S8	1	1	1	1	4.0	High
S9	1	1	1	0.5	3.5	High
S10	1	1	1	0.5	3.5	High
S11	1	1	1	0.5	3.5	High
S20	1	1	1	1	4.0	High
S22	1	1	1	0.5	3.5	High
S25	1	1	1	0.5	3.5	High
S1	1	1	0.5	0.5	3.0	Medium
S12	0.5	1	1	0.5	3.0	Medium
S13	1	0.5	1	0.5	3.0	Medium
S14	1	0	1	1	3.0	Medium
S15	1	0.5	1	0.5	3.0	Medium
S16	1	0.5	1	0.5	3.0	Medium
S18	1	1	0.5	0.5	3.0	Medium
S23	0.5	1	1	0.5	3.0	Medium
S26	1	0.5	1	0.5	2.5	Medium
S5	0.5	0.5	1	0	1.5	Low
S17	0	0.5	1	0.5	2.0	Low
S19	1	0	0.5	0.5	2.0	Low
S21	1	0.5	0.5	0	2.0	Low
S24	1	0	1	0	2.0	Low

Table 11. Score Distribution by Quality Assessment Criterion ($n = 26$)

Score	QA1	QA2	QA3	QA4
Yes (1)	22 (84.6%)	16 (61.5%)	22 (84.6%)	5 (19.2%)
Partly (0.5)	3 (11.5%)	7 (26.9%)	4 (15.4%)	18 (69.2%)
No (0)	1 (3.8%)	3 (11.5%)	0 (0%)	3 (11.5%)
Mean Score	0.90	0.75	0.92	0.54

- Following configurations from other publications without explicit details

QA3 – Validity of Evaluation. This criterion showed the strongest performance across studies. All 26 studies received at least partial scores, with 22 studies (84.6%) receiving full scores. Studies generally employed appropriate metrics (e.g., collision rate, success rate, driving score) and included comparisons with baseline methods or alternative approaches.

QA4 – Transparency of Findings. This criterion showed the greatest weakness across the included studies. Key observations:

- Only 5 studies (19.2%) received full scores for discussing limitations and providing comprehensive results presentation
- The majority (18 studies, 69.2%) received partial scores, typically presenting results clearly but failing to discuss limitations
- 3 studies (11.5%) received no credit, failing to discuss limitations or threats to validity

The mean score for QA4 (0.54) was substantially lower than the other criteria (QA1: 0.90, QA2: 0.75, QA3: 0.92), indicating a significant opportunity for improvement in research transparency within the PPO-for-autonomous-driving community.

7. Implications for Evidence Synthesis

The quality assessment results have important implications for interpreting the synthesized findings:

1) **Overall Acceptable Quality:** With 46.2% of studies rated as high quality and only 19.2% rated as low quality, the evidence base provides a reasonably reliable foundation for synthesis.

2) **Transparency Gap:** The consistent weakness in QA4 suggests that findings should be interpreted with caution, as potential limitations may not be fully reported across the literature.
3) **Replication Challenges:** The variability in QA2 scores indicates that replicating some studies may be difficult due to incomplete methodological reporting.
4) **Evaluation Strength:** The strong performance on QA3 suggests that evaluation methodologies are generally sound, lending credibility to reported performance results.

Acknowledgment

The authors would like to acknowledge the use of AI and LLM for assistance in the literature review. It is important to note that all AI assisted processes were reviewed, verified, and validated by the authors to ensure accuracy, reliability, and adherence to academic standards.

References

[1] SAE, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," 2021, sAE J3016.

[2] X. Dong and M. L. Cappuccio, "Applications of computer vision in autonomous vehicles: Methods, challenges and future directions," *arXiv:2311.09093*, 2024. [Online]. Available: https://arxiv.org/pdf/2311.09093v3

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, Massachusetts: MIT Press, 2018.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[5] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[6] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2016, pp. 1928–1937.

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

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[9] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2016.

[10] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning (ICML)*, 2018, pp. 1587–1596.

[11] 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 (ICML)*, 2018, pp. 1861–1870.

[12] N. Gavenski, F. Meneguzzi, M. Luck, and O. Rodrigues, "A Survey of Imitation Learning Methods, Environments and Metrics," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 37, no. 4, p. 111, aug 2024.

[13] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 627–635.

[14] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[15] B. B. Elallid, N. Benamar, A. S. Hafid, T. eddine Rachidi, and N. Mrani, "A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 7366–7390, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248077878

[16] J. Zhao, Y. Wu, R. Deng, S. Xu, J. Gao, and A. F. Burke, "A survey of autonomous driving from a deep learning perspective," *ACM Computing Surveys*, vol. 57, pp. 1 – 60, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277925267

[17] V. Bharilya and N. Kumar, "Machine learning for autonomous vehicle's trajectory prediction: A comprehensive survey, challenges, and future research directions," *Veh. Commun.*, vol. 46, p. 100733, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259936909

[18] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *International Conference on Learning Representations (ICLR)*, 2016. [Online]. Available: https://arxiv.org/abs/1506.02438

[19] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep Reinforcement Learning for Autonomous Driving: A Survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2022.

[20] Z. Zhu and H. Zhao, "A Survey of Deep RL and IL for Autonomous Driving Policy Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 043–14 065, 2022.

[21] R. Zhao, Y. Li, Y. Fan, F. Gao, M. Tsukada, and Z. Gao, "A Survey on Recent Advancements in Autonomous Driving Using Deep Reinforcement Learning: Applications, Challenges, and Solutions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 19 365–19 398, 2024.

[22] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, vol. 55, no. 12, pp. 2049–2075, 2013.

[23] B. Kitchenham, L. Madeyski, and D. Budgen, "SEGRESS: Software Engineering Guidelines for REporting Secondary Studies," *IEEE Transactions on Software Engineering*, vol. 49, no. 3, pp. 1273–1298, March 2023.

[24] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, p. n71, 2021. [Online]. Available: https://www.bmj.com/content/372/bmj.n71

[25] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University, Technical Report EBSE-2007-01, 2007, version 2.3.

[26] L. Yang, Z. Wang, Y. Zhou, Y. Yang, X. Chen, and B. Xu, "Quality assessment in systematic literature reviews: A software engineering perspective," *Information and Software Technology*, vol. 130, p. 106397, February 2021.

[27] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 202–15 212.

[28] L. M. Schmidt, G. Kontes, A. Plinge, and C. Mutschler, "Can you trust your autonomous car? interpretable and verifiably safe reinforcement learning," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 171–178.

[29] G. C. K. Couto and E. A. Antonelo, "Generative adversarial imitation learning for end-to-end autonomous driving on urban environments," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 1–7.

[30] F. Ye, P. Wang, C.-Y. Chan, and J. Zhang, "Meta reinforcement learning-based lane change strategy for autonomous vehicles," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 223–230.

[31] N. Zhong, J. Chen, Y. Ma, and W. Jiang, "Decision making for driving agent in traffic simulation via adversarial inverse reinforcement learning," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

[32] Y. Wang, H. Ali, X. Dai, K. Wang, and F. Zhu, "Embed trajectory imitation in reinforcement learning: A hybrid method for autonomous vehicle planning," in *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, 2023.

[33] M. Albilani and A. Bouzeghoub, "Guided hierarchical reinforcement learning for safe urban driving," in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2023, pp. 746–753.

[34] J. Shi, T. Zhang, J. Zhan, S. Chen, J. Xin, and N. Zheng, "Efficient lane-changing behavior planning via reinforcement learning with imitation learning initialization," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023.

[35] R. Zhou, H. Cao, J. Huang, X. Song, J. Huang, and Z. Huang, "Hybrid lane change strategy of autonomous vehicles based on SOAR cognitive

architecture and deep reinforcement learning," *Neurocomputing*, vol. 611, p. 128669, 2024.

[36] K. Hao, W. Cui, Y. Luo, L. Xie, Y. Bai, J. Yang, S. Yan, Y. Pan, and Z. Yang, "Adversarial safety-critical scenario generation using naturalistic human driving priors," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 9, pp. 5392–5406, 2024.

[37] R. Zhou, J. Huang, M. Li, H. Li, H. Cao, and X. Song, "Knowledge transfer from simple to complex: A safe and efficient reinforcement learning framework for autonomous driving decision-making," *Advanced Engineering Informatics*, vol. 64, p. 103047, 2025.

[38] J. Chowdhury, V. Shivaraman, S. Sundaram, and P. Sujit, "Graph-based prediction and planning policy network (GP3Net) for scalable self-driving in dynamic environments using deep reinforcement learning," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, 2024, pp. 11 606–11 614.

[39] G. C. K. Couto and E. A. Antonelo, "Hierarchical generative adversarial imitation learning with mid-level input generation for autonomous driving on urban environments," *IEEE Transactions on Intelligent Vehicles*, pp. 1–14, 2024.

[40] X. Ye, F. Tao, A. Mallik, B. Yaman, and L. Ren, "Lord: Large models based opposite reward design for autonomous driving," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 5072–5081.

[41] M. Wen, F. He, Y. Yue, J. Zhang, H. Zhu, and D. Wang, "Target-driven mapless navigation for self-driving car," in *2021 IEEE International Conference on Unmanned Systems (ICUS)*, 2021, pp. 505–511.

[42] F. Konstantinidis, M. Sackmann, U. Hofmann, and C. Stiller, "Graph-based adversarial imitation learning for predicting human driving behavior," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 857–864.

[43] H. Xue, M. Shi, and Q. Jiao, "AWGAIL: Auxiliary human decision-making with Wasserstein GAIL for end-to-end autonomous driving," in *2025 Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS)*, 2025.

[44] R. Rauch, Š. Korečko, and J. Gazda, "Evaluation of proximal policy optimization with extensions in virtual environments of various complexity," in *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2022.

[45] S. Mohammed, A. Argun, and G. Ascheid, "A unified approach to autonomous driving in a high-fidelity simulator using vision-based reinforcement learning," in *2024 IEEE/SICE International Symposium on System Integration (SII)*, 2024, pp. 1093–1098.

[46] S. Dey, S. Pendurkar, G. Sharon, and J. P. Hanna, "A joint imitation-reinforcement learning framework for reduced baseline regret," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3485–3491.

[47] C. Xu, W. Zhao, J. Liu, C. Wang, and C. Lv, "An integrated decision-making framework for highway autonomous driving using combined learning and rule-based algorithm," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 4, pp. 3621–3632, 2022.

[48] X. Sun, M. Zhou, Z. Zhuang, S. Yang, J. Betz, and R. Mangharam, "A benchmark comparison of imitation learning-based control policies for autonomous racing," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023.

[49] J.-H. Choi, J.-T. Hwang, J.-S. Yoo, J.-Y. Kim, B.-J. Kim, and D.-h. Kim, "Enhancing autonomous driving with pre-trained imitation and reinforcement learning," in *2025 International Conference on Electronics, Information, and Communication (ICEIC)*, 2025.

[50] P. Kaur and R. Sobti, "A novel hybrid framework for motion planning in autonomous vehicles using reinforcement and imitation learning," in *2024 International Conference on Computational Intelligence and Network Systems (CINS)*, 2024.

[51] M. Sackmann, H. Bey, U. Hofmann, and J. Thielecke, "Modeling driver behavior using adversarial inverse reinforcement learning," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 1683–1690.

[52] J. Zhu, J. Ortiz, and Y. Sun, "Decoupled deep reinforcement learning with sensor fusion and imitation learning for autonomous driving optimization," in *2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2024, pp. 306–310.