Federated World Models for Privacy-Preserving Collaborative Autonomous Driving in Edge-Vehicle Networks

Akshay Krishnan
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
akshaykrishnan349@colostate.edu

Quentin Salonen
Department of Computer Science, University of Houston, Houston, TX, USA.
quentin.work@uh.edu

Scott Hansen
Department of Computer Science, George Mason University, Fairfax, VA, USA.
scotth@gmu.edu

Matteo J. Ramos
Department of Computer Science, University of North Texas, Denton, TX, USA.
matteo.work@unt.edu

Abstract

The development of safe and robust autonomous driving systems depends critically on the ability to perceive, predict, and plan over complex and dynamic environments. Traditional centralized approaches require the aggregation of vast amounts of sensitive trajectory and visual data from vehicles, raising significant privacy, security, and regulatory concerns. This paper introduces the concept of federated world models, a distributed learning framework that combines the representational power of latent dynamics models with the privacy-preserving properties of federated learning, operating within edge-vehicle networks. Unlike conventional federated learning that focuses on supervised tasks, federated world models enable vehicles to collaboratively learn a shared generative model of the environment—including scene understanding, motion prediction, and counterfactual reasoning—without exposing raw sensor data. We present a comprehensive system architecture where vehicles serve as local clients, edge nodes perform hierarchical aggregation and differential privacy budgeting, and a cloud server maintains a global world model. The paper examines structural trade-offs among model fidelity, communication efficiency, latency constraints, and fairness under heterogeneous driving conditions. It further discusses governance frameworks, policy implications for data sovereignty, infrastructure requirements for low-latency vehicle-to-everything connectivity, and sustainability considerations. Through analysis of current federated learning protocols and world model architectures, we argue that federated world models offer a viable path toward scalable, privacy-compliant, and collaborative autonomous driving, while also highlighting open challenges in adversarial robustness, non-stationary environments, and certification of distributed intelligent systems.

Keywords

federated learning, world models, autonomous driving, privacy-preserving machine learning, edge computing, collaborative perception, socio-technical systems, differential privacy, vehicle-to-everything, distributed intelligence.

1. Introduction

The promise of autonomous driving relies on the ability to continuously sense, interpret, and anticipate the behavior of diverse road agents under constantly shifting conditions. Modern autonomous driving stacks incorporate multiple deep neural networks for perception, prediction, planning, and control, all of which benefit from exposure to large, diverse datasets. However, the collection of raw sensor streams—including camera images, LiDAR point clouds, and GPS trajectories—from a fleet of vehicles raises profound privacy risks, as these data can reveal individuals' locations, driving habits, and even facial images of pedestrians. Regulations such as the General Data Protection Regulation in Europe and the California Consumer Privacy Act impose strict limitations on the centralized storage and processing of personal data, making traditional data-hungry training pipelines increasingly untenable. Federated learning, first proposed by McMahan and colleagues, offers a paradigm in which models are trained across decentralized devices without transferring raw data to a central server [1]. While early works focused on supervised learning tasks such as next-word prediction and image classification, the application of federated learning to autonomous driving introduces unique challenges regarding temporal dynamics, multi-modal sensor fusion, and the need for generative predictive models.

World models, originally developed for reinforcement learning by Ha and Schmidhuber, provide a framework for learning compressed latent representations of an environment that can be used to simulate future states and reason about consequences of actions [7]. Hafner and coauthors extended this concept with deep latent dynamics models that learn to predict pixel-level and reward outcomes from sequences of observations and actions [8][9]. In the context of autonomous driving, a world model can encapsulate the underlying rules of traffic behavior, road geometry, and agent interactions, enabling a vehicle to imagine multiple plausible futures and select safe maneuvers. However, training such a world model requires access to temporally correlated data from many vehicles navigating diverse scenes. Centralizing this data would expose detailed operational profiles, yet sharing only model updates rather than raw data could theoretically preserve privacy. This motivates the proposal of federated world models, where each vehicle locally maintains a latent dynamics model, periodically updates a shared global model via an edge infrastructure, and never reveals its raw observations.

Edge computing, as surveyed by Shi and colleagues, brings computation and storage resources closer to data sources, reducing latency and bandwidth usage [6]. In a federated world model architecture, edge nodes located at roadside units or local aggregation points can perform intermediate model averaging, apply differential privacy mechanisms, and manage client selection. Vehicles communicate with these edge nodes over vehicle-to-everything (V2X) links, which are themselves evolving with 5G and future 6G networks to support ultra-reliable low-latency communication. The synthesis of federated learning, world models, and edge computing creates a system-level solution that addresses both the technical demands of autonomous driving and the societal imperatives of privacy and data governance. The remainder of this paper unpacks the architectural, trade-off, governance, and deployment dimensions of this approach.

2. Background and Related Work

Federated learning has matured from a theoretical concept to a practical framework for distributed machine learning, with numerous variants addressing communication efficiency, heterogeneous data distributions, and security [3][19]. Konecny and colleagues introduced strategies to reduce communication rounds through compressed updates and structured random rotations [2]. Subsequently, secure aggregation protocols were developed to ensure that the server learns only the aggregated model and not individual client updates, as demonstrated by Bonawitz and colleagues [5]. Differential privacy, formalized by Dwork and Roth, provides a rigorous framework for bounding information leakage from query outputs, and has been integrated into federated learning workflows by adding calibrated noise to client updates or aggregated models [4]. Despite these advances, federated learning applied to high-dimensional, temporally structured tasks such as video prediction or trajectory forecasting remains challenging because the local data distributions are non-independent and identically distributed (non-IID) across vehicles, and the temporal dependencies require careful handling of sequence lengths and state initialization.

World models have been adopted in autonomous driving research as a means to compress high-dimensional sensory input into a compact latent state that captures scene dynamics. Hu and colleagues demonstrated model-based imitation learning for urban driving, leveraging a learned dynamics model to generate additional training trajectories [10]. Meanwhile, end-to-end driving systems, as pioneered by Bojarski and colleagues, learn direct mappings from camera pixels to steering commands but lack explicit predictive reasoning and are vulnerable to distribution shift [12]. Chen and colleagues introduced affordance-based direct perception, which predicts intermediate representations such as road curvature and distance to lead vehicles, bridging the gap between end-to-end learning and modular systems [11]. World models offer a middle ground: they learn a generative model of the environment that can be used both for planning and for data augmentation, all while operating in a latent space that is more efficient for communication. However, the training of world models in a distributed, privacy-preserving manner has received limited attention, with most existing federated learning research focusing on discriminative tasks.

A recent survey by Liang and colleagues specifically examined federated learning for autonomous driving, identifying challenges such as client heterogeneity, communication constraints, and safety verification [13]. They noted that most prior work applied federated learning to perception tasks like object detection, leaving prediction and planning largely unexplored. Edge intelligence for autonomous driving has been surveyed by Huang and colleagues, who highlighted the role of edge servers in reducing inference latency and enabling real-time collaborative perception [20]. Yet, the combination of federated learning with generative world models on edge-vehicle networks has not been systematically studied. This paper attempts to fill that gap by proposing a coherent architecture and analyzing the system-level implications.

3. System Architecture of Federated World Models

The proposed system architecture comprises three tiers: the vehicle tier, the edge tier, and the cloud tier. Each vehicle operates as a federated client, running a local instance of a world model that consists of an encoder, a latent dynamics module, and a decoder. The encoder maps high-dimensional sensor observations—such as camera images, LiDAR point clouds, and radar data—into a compact latent representation. The dynamics module predicts future latent states given current state and planned actions, typically using a recurrent neural network or a transformer-based sequence model. The decoder reconstructs observations, allows for

imagining future scenes, and can also predict auxiliary quantities such as occupancy grids or drivable areas. During local training, the vehicle samples sequences of observations and actions from its own driving logs and updates its local model parameters to minimize a combination of reconstruction loss, prediction loss, and possibly a reward or safety cost.

Periodically, each vehicle communicates its model update to a nearby edge node rather than directly to a central server. Edge nodes are deployed at intersections, highway gantries, or cellular base stations, and they are responsible for aggregating updates from vehicles within a geographic region. The edge node performs secure aggregation, for instance using the protocol of Bonawitz and colleagues, to compute the sum or average of the encrypted updates without revealing individual contributions [5]. Additionally, the edge node can apply differential privacy by adding Gaussian or Laplacian noise to the aggregated update, with noise scale calibrated to a pre-defined privacy budget that is tracked across rounds. The edge node then forwards the noisy aggregated update to the cloud server. The cloud server maintains the global world model, which is updated by averaging the aggregated updates from all edge nodes. This hierarchical aggregation scheme reduces communication overhead compared to a fully centralized topology, as vehicles communicate only with nearby edge nodes, and edge nodes compress multiple updates into a single message to the cloud.

A critical design choice is the synchronization frequency. Synchronous federated learning requires all vehicles to complete a round of local training before aggregation can proceed, which can lead to stragglers and delays in the presence of heterogeneous compute capabilities and network conditions. Asynchronous aggregation, on the other hand, allows updates to be applied immediately, but can cause model staleness and convergence issues, especially when the data generation process is non-stationary. For world models, which require coherent temporal sequences for effective training, partial synchronization schemes such as semi-asynchronous protocols or temporally weighted averaging may be more appropriate. We also note that vehicles may disconnect and reconnect as they move through different edge coverage zones; the system must support client unavailability and state transfer across edge nodes, akin to a handover mechanism. Edge nodes themselves may need to share lightweight signatures of the local world model state to ensure smooth continuation when a vehicle crosses a boundary, raising additional privacy considerations regarding the transfer of intermediate latent representations.

4. Structural Trade-offs and Design Considerations

The federated world model architecture involves several inherent trade-offs that must be carefully balanced. Model accuracy, privacy, communication efficiency, latency, and fairness are often in tension. For instance, increasing the complexity of the world model improves predictive fidelity but also increases the size of model updates transmitted over the network, consuming bandwidth and energy. Compression techniques such as gradient sparsification, quantization, and low-rank factorization have been proposed in federated learning literature to reduce communication costs [2]. However, world models contain recurrent components that may be less amenable to aggressive compression because small errors in latent dynamics can accumulate over long rollouts. A trade-off therefore exists between communication efficiency and the generative quality of the learned world model.

Privacy protection introduces its own set of design decisions. Differential privacy ensures that the presence or absence of any individual vehicle's data cannot be inferred from the released model, but it requires adding noise that degrades model accuracy. The noise magnitude grows with the number of training rounds, meaning that privacy budgets must be carefully managed.

In a federated world model, the local training process may involve many gradient steps per round, and the sensitivity of the model update depends on the bounds of the loss function. Practitioners must decide whether to apply differential privacy at the client level (adding noise to local updates before sending), at the edge level (adding noise after aggregation), or both. Client-level differential privacy provides stronger guarantees because an adversary cannot distinguish which client contributed, even if edge nodes are compromised. However, it reduces the effective signal-to-noise ratio, especially when the number of participating clients in a region is small. Secure aggregation prevents the edge or cloud from seeing individual updates, but does not protect against inference attacks on the final model—a combination of both techniques is often recommended [5][17].

Latency is a critical constraint for autonomous driving applications, where control decisions must be made in milliseconds. Federated world model training occurs offline or periodically, but the real-time inference of the world model on the vehicle must meet strict deadlines. The compressed latent representation used for inference is computationally lightweight, so the online latency is acceptable. However, the aggregation process—particularly synchronous rounds—can introduce delays that affect the freshness of the global model. An outdated world model may not reflect recent changes in the environment (e.g., new construction zones or temporary road closures). Adaptive synchronization schedules that trigger retraining only when a significant distribution shift is detected, or that use meta-learning to quickly adapt to new scenes, could alleviate this issue [14]. Fairness across vehicles arises due to non-IID data distributions: vehicles operating in urban centers encounter dense traffic and varied interactions, while rural vehicles may experience long stretches of highway with repetitive patterns. A federated world model that aggregates updates uniformly may become biased toward the more common data type, potentially degrading performance in minority scenarios. Strategies such as weighted aggregation based on data diversity, fairness constraints, or personalized model branches for different driving contexts can help mitigate this imbalance.

5. Governance, Infrastructure, and Deployment

Deploying federated world models at scale requires a robust governance framework that addresses data sovereignty, liability, and certification. Since the vehicles are owned by individuals or fleets, and the edge infrastructure may be operated by municipalities, telecommunications companies, or third-party service providers, clear agreements must define who controls the model, what data is used, and how privacy is guaranteed. Regulations such as GDPR require that personal data be processed only with explicit consent and that individuals can request deletion of their data; in a federated world model, the raw data never leaves the vehicle, but the model itself may encode information about the training data. Differential privacy provides a legal pathway by ensuring that removal of any single vehicle's data does not change the model's outputs beyond a bounded amount, but the level of epsilon must be chosen to satisfy regulatory thresholds—a topic of ongoing debate.

Infrastructure readiness is another major hurdle. Edge nodes must be deployed with sufficient computational capacity to handle multiple aggregation rounds per second, especially during peak traffic hours. The communication network must support high-bandwidth, low-latency V2X links. Existing 5G networks can achieve latencies below ten milliseconds, but wide-area coverage for autonomous driving may require densification of base stations and the integration of dedicated short-range communication (DSRC) or cellular V2X. Energy consumption of the overall system is a growing concern: training deep world models on vehicle-level hardware consumes battery power, and the edge nodes and cloud servers require

significant electricity. Federated learning can reduce energy compared to centralizing all data, but the communication and computation overhead must be optimized. Sustainable deployment may involve scheduling training during off-peak hours or when vehicles are charging, and using energy-efficient hardware accelerators.

Finally, safety certification of a federated world model poses unprecedented challenges. Unlike a deterministic control algorithm, a neural network world model is a black box whose behavior cannot be fully verified. Acceptance by automotive regulators, such as the National Highway Traffic Safety Administration or the European Union's type-approval authorities, will require demonstrating that the distributed training process does not introduce hidden biases or failures that could lead to accidents. Techniques such as formal verification of latent space properties, adversarial testing with counterfactual scenarios, and runtime monitoring of prediction uncertainty may become part of the certification process. The fact that the model is continuously updated via federated learning introduces a moving target; regulators may require a snapshot of the model at deployment time and impose constraints on how much it can change online.

6. Future Directions and Conclusion

The concept of federated world models opens several promising research avenues. One direction is the development of meta-learning algorithms that allow the global world model to quickly adapt to new driving domains with minimal local data, reducing the number of communication rounds required. Another is the integration of multimodal world models that jointly reason over vision, language (e.g., traffic signs), and high-definition maps, while preserving privacy through federated learning across different sensor types. Cross-domain transfer between autonomous driving and other robotic applications, such as last-mile delivery or warehouse logistics, could leverage a common world model backbone trained in a federated manner. The governance and policy aspects of distributed intelligence demand deeper collaboration between computer scientists, legal scholars, and transportation authorities to establish standards for data sharing, liability distribution, and fairness auditing.

In conclusion, federated world models represent a synthesis of two powerful paradigms—federated learning and generative world models—anchored in an edge-vehicle network infrastructure. By enabling vehicles to collaboratively learn a shared understanding of driving environments without exposing raw data, this approach addresses critical privacy and regulatory concerns while potentially improving the safety and robustness of autonomous driving systems. The paper has outlined the architectural components, examined the key trade-offs among accuracy, privacy, communication, latency, and fairness, and discussed the governance, infrastructure, and certification challenges that must be surmounted for real-world deployment. As autonomous driving technology progresses from controlled testbeds to widespread adoption, federated world models offer a principled pathway that respects individual privacy while harnessing collective intelligence.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54, 1273-1282.

2. Konecny, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

3. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

4. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

5. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175-1191.

6. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637-646.

7. Ha, D., & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.

8. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2020). Learning latent dynamics for planning from pixels. Proceedings of the 36th International Conference on Machine Learning, 97, 2555-2565.

9. Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2021). Mastering Atari with discrete world models. arXiv preprint arXiv:2010.02193.

10. Hu, A., Corrado, G., Griffiths, N., Murez, Z., Gurau, C., Yeo, H., ... & Rao, K. (2023). Model-based imitation learning for urban driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

11. Chen, C., Seff, A., Kornhauser, A., & Xiao, J. (2015). DeepDriving: Learning affordance for direct perception in autonomous driving. Proceedings of the IEEE International Conference on Computer Vision, 2722-2730.

12. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... & Zhang, H. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

13. Liang, P. P., Cheng, Y., & Salakhutdinov, R. (2022). Federated learning for autonomous driving: A survey. arXiv preprint arXiv:2205.13668.

14. Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., & others. (2021). A field guide to federated optimization. arXiv preprint arXiv:2107.06917.

15. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775.

16. Xiong, Z., Ye, X., Yaman, B., Cheng, S., Lu, Y., Luo, J., ... & Ren, L. (2026). UniDrive-WM: Unified Understanding, Planning and Generation World Model For Autonomous Driving. arXiv preprint arXiv:2601.04453.

17. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in Neural Information Processing Systems, 32.

18. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1-19.

19. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210.

20. Huang, Y., Chen, Y., & Zhao, R. (2023). Edge intelligence for autonomous driving: A survey. IEEE Transactions on Intelligent Vehicles, 8(2), 1128-1144.