Vision-Language World Models for Risk-Aware Decision Making in Autonomous Driving

Aarav Yadav
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
aarav1993@oregonstate.edu

Matteo R. Lindgren
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
matteolindgren83@binghamton.edu

Lars Murray
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
lars.work@buffalo.edu

Edwin M. Peters
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
helloedwin@unh.edu

Abstract

Autonomous driving has increasingly moved from modular perception-planning pipelines toward learned systems that must operate under profound uncertainty. Vision-language world models offer a promising mechanism for anticipating future states, interpreting traffic situations through natural language, and supporting risk-aware decisions. This paper examines their role from a systems perspective, focusing on architectural trade-offs, data infrastructure, operational deployment, and governance rather than narrowly algorithmic improvements. We argue that unifying visual understanding, language-grounded reasoning, and predictive generation changes the nature of risk assessment because the model can represent counterfactual futures, contextualize ambiguous observations, and explain potential conflicts in human-interpretable terms. However, these capabilities introduce structural challenges including calibration of multimodal uncertainty, integration with existing safety cases, and dependence on large heterogeneous datasets. We discuss how reinforcement learning, generative modeling, and instruction tuning can be coupled to improve decision quality under uncertainty while preserving oversight. The analysis covers robustness against distribution shift and adversarial perturbation, fairness across road users and environments, sustainability of large-scale training, and the policy implications of deploying such models in public infrastructure. We conclude that vision-language world models should be treated as socio-technical control systems rather than isolated perception modules. Their adoption requires continuous monitoring, standardized evaluation across risk dimensions, transparent documentation, and regulatory alignment with existing vehicle safety frameworks. The paper contributes a cross-disciplinary synthesis of architecture, governance, and deployment concerns intended to inform researchers, engineers, and policymakers.

Keywords

autonomous driving; vision-language models; world models; risk-aware decision making; socio-technical systems.

1. Introduction

The evolution of autonomous driving has been shaped by deep learning [1] and attention-based neural architectures [2]. Early modular systems separated detection, tracking, prediction, and planning into distinct components with hand-designed interfaces. This modularity simplified debugging but often led to error accumulation across subsystem boundaries. Learned representations and attention mechanisms made it possible to integrate multimodal sensor streams more tightly, improving scalability and adaptation to complex urban scenes. As a result, the field has shifted toward architectures that can reason over scenes in a more holistic manner. However, this shift also changes how risk must be understood because errors no longer remain confined to a single module.

Foundation models have accelerated this transformation [3]. Vision-language models align visual features with natural language descriptions, allowing semantic queries over visual scenes [4]. Large language models provide few-shot reasoning and contextual interpretation across many domains [5]. In autonomous driving, these capabilities can be used to encode traffic rules, road signs, map context, and driver intent. Natural language offers a flexible interface for human operators, regulators, and downstream planning modules. At the same time, language introduces new failure modes related to semantic ambiguity, overconfidence, and the possibility of plausible but incorrect descriptions of visual events.

Risk-aware decision making further requires the ability to imagine possible future states that have not yet occurred. Generative world models provide a computational substrate for this imagination. Diffusion models generate high-quality future scenes by iterative denoising [6], while generative adversarial networks enable realistic multimodal synthesis through adversarial training [7]. These generative mechanisms allow a system to represent multiple plausible futures conditioned on different actions and environmental factors. When coupled with language, such futures can be expressed as interpretable scenarios with semantic labels, supporting both automated risk evaluation and human review. The integration of vision, language, and generation therefore represents a meaningful architectural transition for autonomous driving.

Despite these advances, the system-level implications of vision-language world models remain underexamined. It is not enough to demonstrate improved trajectory prediction or image generation. The deployment of such models in real vehicles raises questions about uncertainty calibration, data governance, safety assurance, fairness, sustainability, and public accountability. This paper analyzes these concerns from an interdisciplinary systems perspective. We discuss how architectural choices interact with operational constraints and regulatory expectations. The emphasis is not on a single benchmark or model, but on the structural trade-offs that shape the safe and responsible use of vision-language world models in autonomous driving.

2. Background and Related Work

Sequential decision making under uncertainty has been studied extensively through deep reinforcement learning. Foundational work demonstrated that convolutional networks can learn control policies directly from high-dimensional visual inputs [8]. Subsequent systems combined learned value functions with search to achieve superhuman performance in strategic games [9]. Proximal policy optimization later provided a more stable approach for

continuous policy learning [10]. These methods are relevant because autonomous driving is fundamentally a sequential control problem with delayed consequences. However, reinforcement learning alone often lacks semantic interpretability, making it difficult to explain why a particular action was selected in a safety-critical situation.

Autonomous driving has also explored end-to-end behavior cloning. Early systems mapped camera images directly to steering commands using deep convolutional networks [11]. This approach demonstrated that low-level driving could be learned from human demonstrations, but it struggled with complex urban interactions and rare events. Later studies exposed the limitations of behavior cloning by showing that errors can compound over time and that purely reactive policies may fail to capture the long-horizon structure of traffic negotiation [12]. These findings motivate predictive models that can anticipate future states rather than merely react to current observations.

Uncertainty quantification is essential for risk-aware deployment. Bayesian deep learning methods distinguish between epistemic uncertainty, which arises from limited knowledge, and aleatoric uncertainty, which arises from irreducible observation noise [13]. In driving, epistemic uncertainty may appear in unfamiliar environments, while aleatoric uncertainty may come from sensor noise or unpredictable pedestrian motion. Robustness is further challenged by adversarial perturbations that can be physically printed and placed in the environment to alter visual classification [14]. A reliable world model must therefore remain calibrated under both natural distribution shift and intentional disruption. These concerns extend beyond accuracy and require explicit system-level safeguards.

3. System Architecture for Vision-Language World Models

A system-level architecture for vision-language world models typically includes a multimodal encoder, a temporal predictive core, a language interface, and a planning head. The multimodal encoder consumes camera, lidar, radar, and map data, aligning these streams into a shared latent representation. The temporal predictive core generates future latent states conditioned on candidate actions and environmental context. The language interface maps visual and temporal representations into natural language descriptions, while also accepting textual queries about rules, goals, and constraints. The planning head evaluates imagined futures and selects actions according to a risk-aware objective. This architecture differs from traditional pipelines because the same latent representation supports perception, prediction, planning, and communication.

Datasets such as nuScenes [15] and Waymo Open Dataset [16] have enabled large-scale training of multimodal driving models. They provide synchronized camera, lidar, radar, map, trajectory, and annotation data across diverse urban scenes. Such datasets are critical for learning the statistical regularities of traffic, including interactions among vehicles, pedestrians, cyclists, and infrastructure. However, they also encode geographic and operational biases. Most large open datasets are collected in a limited set of cities under particular weather and lighting conditions. A world model trained on these distributions may imagine futures that are plausible for those locations but misleading elsewhere. Data coverage therefore directly shapes the counterfactual space available to the model.

Recent unified architectures integrate understanding, planning, and generation within a single world model framework [17]. Instead of maintaining separate modules for detection, tracking, motion forecasting, and planning, these architectures generate multimodal future scenes conditioned on language and planned action. This integration reduces interface error and

allows the planner to query the same predictive representation used for scene understanding. However, it also concentrates epistemic risk in one model family. When a single model is responsible for perception, prediction, and planning, failures can propagate without clear component boundaries. Verification becomes more difficult because traditional safety cases assume that independent subsystems can be analyzed separately. Unified world models therefore require new forms of assurance that account for entangled representations.

4. Risk-Aware Decision Making and Uncertainty Representation

Risk awareness requires a model to distinguish ordinary variability from dangerous tail events. A vision-language world model can generate multiple future trajectories, associate them with semantic descriptions, and evaluate outcomes under different actions. This supports counterfactual reasoning, which is central to safe driving. For example, the system may imagine a future in which a pedestrian steps into the road under low visibility, and then compare that future with one in which the pedestrian remains on the sidewalk. The ability to generate and describe such alternatives improves risk assessment beyond passive prediction. Yet generated futures may not cover long-tail failures, especially when the model has never observed similar situations during training.

Safety problems in AI systems include negative side effects, reward hacking, safe exploration, and distributional shift [18]. In autonomous driving, these problems appear as interactions with vulnerable road users, exploitation of simulator artifacts, and poor generalization beyond the operational design domain. A model trained on dense urban data may perform poorly in rural areas, construction zones, or adverse weather. Risk-aware deployment therefore requires explicit monitoring of distributional mismatch and intervention mechanisms. The vehicle must recognize when its world model is no longer reliable and transition to a conservative fallback policy or request human assistance.

The semantic layer of a vision-language model can also represent traffic rules, road signs, and driver intent in natural language. This enables the system to reason about violations and ambiguous right-of-way. For instance, a model might explain that a vehicle is approaching a yield sign with limited visibility. However, language-grounded risk is not automatically calibrated. The model may assign high linguistic confidence to a wrong visual interpretation. Calibration across image, language, and policy remains an open system engineering challenge. Risk-aware decision making must therefore combine calibrated uncertainty estimates with semantic explanations, rather than treating language fluency as evidence of correctness.

5. Training Data Infrastructure and Model Alignment

Training vision-language world models requires extensive data infrastructure. Data must be curated, anonymized, labeled, and balanced across operational contexts. Large pretraining datasets may contain near-duplicates, annotation errors, and skewed distributions. The resulting model can propagate geographic and social biases. Alignment with human preferences through reinforcement learning from human feedback can improve instruction following and safety behavior [19]. In driving, such alignment might teach the model to prefer conservative gap acceptance or to avoid abrupt maneuvers. However, human feedback for driving is expensive and may reflect inconsistent preferences across annotators and regions.

Chain-of-thought prompting and related reasoning methods can make model decisions more transparent by generating intermediate textual explanations [20]. This is useful for auditing and for aligning risk logic with traffic norms. A model might explain that it slowed down because the pedestrian was visually occluded and could enter the crosswalk. Yet explicit

reasoning may not fully reflect actual internal computations. It can also be influenced by spurious linguistic priors. Explanations should therefore be treated as auxiliary signals that support oversight, not as complete accounts of model behavior. They are valuable for human interpretability, but they do not replace formal verification or empirical safety testing.

Data governance must include provenance, consent, and scenario coverage. Real-world data cannot be collected exhaustively for rare fatal events. Synthetic data and world models themselves may augment training sets, but this introduces recursive risk because models trained on model-generated futures can amplify systematic blind spots. Infrastructure should track data lineage, corruption, and evaluation splits to prevent leakage and overoptimistic safety claims. Training pipelines should also preserve metadata about weather, geography, sensor configuration, and traffic density. Such metadata allows operators to define the operational design domain and to monitor how well the deployed model matches its training distribution.

6. Deployment, Verification, and Operational Governance

Deployment of world models in autonomous driving shifts verification from component-level to system-level analysis. Traditional safety cases identify hazards and assign mitigation across sensors, planners, and actuators. A unified model blurs these boundaries because perception, prediction, and planning share latent representations. Assurance must address representational completeness, runtime monitoring, and graceful degradation. Operational governance includes geofencing, speed restrictions, weather limits, and remote human oversight. These restrictions reduce exposure to situations where the model is likely to fail. They also create a controlled environment for collecting operational data and updating the model.

Continuous monitoring can compare predicted futures with observed outcomes. Calibration metrics such as mismatch between predicted and actual trajectories should be tracked across regions, times of day, and demographic groups. Governance should also include event reporting, post-incident analysis, and model update protocols. Each update may alter risk in subtle ways, so staged rollout and shadow mode evaluation are important. In shadow mode, the model observes and predicts without controlling the vehicle, allowing comparison with human behavior before deployment. This reduces the risk of silent regression and provides evidence for safety claims.

Human factors are equally important. Driver attention, handover latency, and trust in system explanations influence risk. A vision-language model that produces plausible but incorrect descriptions can undermine trust or induce overreliance. System design must manage the interaction between model confidence, semantic output, and human oversight. For example, a vehicle should not present a detailed explanation of a predicted future if that explanation gives a false sense of certainty. The interface should communicate uncertainty in a way that supports appropriate supervisory control. This requires collaboration across engineering, human factors, and regulatory bodies.

7. Robustness, Fairness, and Sustainability

Robustness under distribution shift and adversarial perturbation is critical for autonomous driving. Physical adversarial patterns can alter model perception and propagate into planning [14]. World models may generate plausible futures that fail under corrupted inputs. Robustness engineering includes input sanitization, uncertainty estimation [13], contrastive training, and runtime anomaly detection. However, no single technique is sufficient. The operational design domain must be specified and enforced, and fallback systems should not

depend on the same visual representations that may be compromised. Defense in depth remains necessary even when using a unified world model.

Fairness concerns arise because autonomous systems may perform differently for pedestrians, cyclists, children, wheelchair users, and regions with different traffic norms. Datasets dominated by certain cities may underrepresent informal settlements, snow, fog, and diverse sensor configurations. A language interface can help specify fairness-relevant context, but it may also introduce textual priors that disadvantage groups whose visual appearance or movement patterns are systematically underrepresented. Equity audits should be part of model evaluation. These audits should measure not only average safety rates but also differential outcomes across road user categories and environmental conditions. Fairness in this context is not merely an abstract ideal; it is directly tied to whose safety is prioritized in ambiguous situations.

Sustainability includes energy consumption, carbon emissions, and hardware lifecycle. Large vision-language models require significant compute during pretraining and fine-tuning. The environmental cost must be weighed against safety benefits. Sparse inference, model distillation, and region-specific adaptation can reduce footprint. However, smaller models may lose counterfactual richness and linguistic reasoning ability. Sustainability therefore is not only an operational concern but also a risk trade-off. Organizations that deploy such systems should report training and inference energy estimates alongside safety metrics. This creates accountability for the broader resource implications of autonomous driving systems.

8. Policy and Socio-Technical Implications

Governance of autonomous driving world models must integrate technical standards, liability, and public accountability. Existing vehicle safety frameworks may not fully address probabilistic language-grounded systems. Policy should require safety cases that specify operational design domains, uncertainty budgets, and human oversight. Independent audits of data, model behavior, and incident data can improve accountability. Regulators may also require simulation-based testing across standardized scenarios, but such testing must be complemented by real-world validation and continuous monitoring. The goal is not to eliminate all risk, which is impossible, but to make risk legible and manageable within publicly accepted boundaries.

Global AI ethics guidelines emphasize transparency, justice, non-maleficence, responsibility, and privacy [21]. These principles are directly relevant to autonomous driving because decisions made by vehicles affect life and safety. Ethical frameworks such as AI4People argue for a society-oriented approach that balances innovation with protection [22]. For world models, this implies documenting training data provenance, explaining risk trade-offs, and ensuring that affected communities have representation in scenario selection and evaluation. Public involvement should not be limited to post hoc consultation; it should shape the priorities embedded in model alignment and deployment policies.

Public trust depends on demonstrable safety outcomes, not only technical accuracy. Regulators may require staged deployment, independent testing, and algorithmic impact assessments. Cities and transportation agencies should have access to model evaluation results and incident data. This kind of socio-technical governance is essential because world models are not merely tools but embedded components of public mobility infrastructure. Their behavior influences how streets are experienced, who can move safely, and how the

public understands automated decision making. Treating them as sociotechnical systems enables a more realistic assessment of their risks and benefits.

9. Conclusion

Vision-language world models represent a significant architectural direction for risk-aware decision making in autonomous driving. By unifying visual understanding, natural language, and predictive generation, these models can represent counterfactual futures and support interpretable planning. This integration may reduce interface errors and improve the ability to reason about ambiguous traffic situations. At the same time, it concentrates epistemic risk, complicates verification, and raises new challenges for uncertainty calibration, data governance, robustness, fairness, and sustainability. The systems perspective developed in this paper suggests that these models should not be evaluated solely on trajectory prediction accuracy or generation quality. They must be assessed as part of a broader socio-technical control loop that includes human oversight, organizational protocols, and regulatory frameworks.

Future work should develop standardized risk metrics that capture multimodal uncertainty and long-tail safety outcomes. Data infrastructure must be designed to document provenance, scenario coverage, and bias. Governance mechanisms should require staged deployment, continuous monitoring, and independent audits. The policy environment must evolve alongside technical capabilities, ensuring that safety cases remain meaningful for unified model architectures. The successful deployment of vision-language world models will depend less on any single architectural innovation than on the ability of institutions to manage the structural trade-offs described here. A deliberate and interdisciplinary approach is necessary to realize the safety and equity benefits of these systems while containing their risks.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748–8763.
5. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
6. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

8. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.

9. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.

10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

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

12. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). Exploring limitations of behavior cloning for autonomous driving. IEEE International Conference on Robotics and Automation (ICRA), 5185–5190.

13. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584.

14. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1625–1634.

15. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11621–11631.

16. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., ... & Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo open dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2446–2454.

17. 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.

18. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

20. Wei, J., Wang,X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

21. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

22. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society:

Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.