Physics-Coherent Autonomous Driving Scene Forecasting through Joint Image-to-Video Generation and 3D World Alignment

Leo Powell
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
leopowell350@uab.edu

Samuel J. Lawson
School of Computing, Clemson University, Clemson, SC, USA.
contactsamuel@clemson.edu

Jase Ceaman
Department of Computer Science, University of North Texas, Denton, TX, USA.
jose.coleman@unt.edu

Abstract

The reliable forecasting of dynamic traffic scenes remains a foundational challenge for the safe deployment of autonomous driving systems. Existing prediction paradigms frequently operate either in the abstract latent space of bird's-eye-view representations or through purely photorealistic video generation without explicit grounding in three-dimensional physical constraints, leading to predictions that violate fundamental laws of motion, occlusion geometry, and object permanence. This paper presents a comprehensive system-level investigation into a novel architectural paradigm that jointly optimizes image-to-video generative models and explicit 3D world alignment modules to produce physics-coherent future scene forecasts. By integrating differentiable volumetric rendering, neural radiance fields, and diffusion-based video synthesis within a shared representation framework, the approach enforces consistency across appearance, geometry, and dynamics. We examine the structural trade-offs involved in balancing generative expressiveness with deterministic physical reasoning, analyze the governance and safety implications of deploying such forecasting models in high-stakes environments, and explore the infrastructure and sustainability challenges inherent in training and serving large-scale generative world models. The discussion extends to robustness under distributional shift, fairness across diverse road user categories, and the evolving regulatory landscape surrounding learned forecasting components. Through deep interdisciplinary analysis, this paper articulates the systemic requirements, risks, and opportunities for joint generation-alignment architectures as a pathway toward trustworthy and physically grounded autonomous scene intelligence.

Keywords

autonomous driving, scene forecasting, image-to-video generation, 3D world alignment, physics coherence, generative models, neural rendering, safety assurance, socio-technical systems.

1. Introduction

The capacity to anticipate the evolution of complex driving environments over multiple seconds into the future is a cornerstone of safe and socially compliant autonomous vehicle behavior. Modern autonomous stacks decompose the forecasting problem into a constellation of sub-tasks, including motion prediction of individual agents, semantic occupancy estimation, and sensor simulation, each addressed by specialized neural architectures. However, this decomposition frequently sacrifices holistic physical consistency, producing forecasts that may be individually accurate yet collectively impossible when projected onto the shared three-dimensional geometry of the world. For instance, predicted bounding boxes may overlap in physically forbidden configurations, or synthesized future video frames may fail to respect occlusion relationships defined by static scene structure. Addressing these inconsistencies demands a fundamental rethinking of how generative forecast models interface with explicit 3D world representations.

Recent advances in large-scale generative modeling, particularly diffusion-based and autoregressive transformers, have demonstrated remarkable fidelity in photorealistic video synthesis, as evidenced by progress in text-conditioned and image-conditioned video generation [5]. Parallel developments in neural radiance fields and volumetric rendering have enabled differentiable pipelines capable of learning continuous 3D scene representations from sparse posed imagery [3]. The convergence of these two threads opens the possibility of constructing forecasting systems that produce future imagery while simultaneously inferring and maintaining a persistent, metric-scale 3D world state. Such systems hold the promise of inherently satisfying physical constraints such as object permanence, rigid-body motion, and view-consistent appearance, provided the generative and geometric modules are optimized under a unified alignment objective.

The challenge of physics-coherent forecasting, however, extends beyond technical architecture. It implicates systemic questions of trustworthiness, robustness under long-tail scenarios, and equity of prediction quality across diverse environmental conditions and road user populations [1]. Autonomous vehicles are sociotechnical systems: their deployment context includes heterogeneous traffic cultures, underrepresented vulnerable road users, and adversarial weather phenomena that stress learned forecasting models. The joint image-to-video and 3D world alignment paradigm is thus not merely a model design problem but a larger systems integration problem, touching on sensor configuration, compute orchestration, online adaptation, regulatory certification, and lifecycle environmental impact.

This paper provides an extended interdisciplinary examination of the joint generation and alignment forecasting framework from a systems perspective. We analyze the architectural design space, explicating the trade-offs between end-to-end learned generation, explicit geometric reasoning, and hybrid modular designs. We discuss how physics coherence can be operationally defined, evaluated, and enforced through shared latent representations and self-supervised losses that couple video prediction with depth, flow, and occupancy. Further, we assess the governance, fairness, and sustainability dimensions that will determine the real-world viability of these forecasting technologies. The goal is to articulate a holistic research agenda and a system-level critique, rather than to propose a singular algorithmic novelty, thereby contributing to the design of autonomous systems that are simultaneously more capable, more accountable, and more aligned with physical reality.

2. System Architecture and Structural Trade-offs

The design of a joint generation-alignment forecasting system necessitates a careful architectural mediation between two traditionally separate families of models: generative

video models that synthesize high-dimensional pixel outputs and geometric alignment modules that reason about metric 3D structure. A straightforward two-stage pipeline, where a video generation model produces future frames that are subsequently lifted to 3D via a separate structure-from-motion or monocular depth network, suffers from error compounding and the absence of mutual constraints. A more integrated architecture instead embeds both capabilities within a shared latent representation that is simultaneously decoded into future images and into explicit 3D geometry, such as voxel grids, triplanes, or point clouds, enabling bidirectional consistency enforcement.

One prominent design pattern involves equipping a diffusion-based video generator with auxiliary heads that predict a 3D occupancy field and a dense depth map for each forecasted timestep. The training objective then jointly minimizes a photometric reconstruction loss and a geometric alignment loss computed by comparing the renderings of the predicted 3D occupancy with the generated frames under known camera intrinsics. This coupling transforms the video generator from an unconditionally imaginative synthesizer into a world-state-conditional forecaster, where visual plausibility is verified against geometric feasibility. Such architectures build upon the insight that neural radiance fields can serve as a differentiable inductive bias for enforcing view consistency and spatial coherence [3]. The trade-off lies in the increased computational burden and the risk of geometric representations limiting the expressiveness of fine-grained texture or transient lighting effects that do not project cleanly onto static 3D structures.

Alternatively, multi-stream architectures process explicit geometric cues, such as lidar point clouds or multi-view depth estimates, through a separate 3D encoder that injects structural priors into the video generation backbone via cross-attention mechanisms. These geometric features act as conditioning signals that guide the frame synthesis toward respecting scene layout, occlusions, and metric distances. The bird’s-eye-view (BEV) representation, commonly used in behavior prediction and sensor fusion, provides a natural intermediate space where spatiotemporal dynamics and geometry can be jointly reasoned before being decoded into perspective imagery [6]. The key system-level decision here concerns the granularity of geometric supervision: enforcing pixel-wise depth consistency may smooth over thin structures like poles and pedestrians, while sparse lidar supervision leaves open ambiguities in textureless regions. The calibration of this trade-off directly impacts the model’s robustness in scenarios featuring construction zones, irregular road edges, or partially occluded vulnerable road users.

Another dimension of architectural choice involves the temporal horizon and granularity of forecasting. A system that generates a dense sequence of 30 Hz video frames while maintaining a consistent 3D world model imposes severe memory and latency constraints, particularly for real-time onboard deployment. Hierarchical architectures mitigate this by first forecasting a temporally coarser latent future state that captures keyframe geometry and semantics, and then upsampling in time via a lightweight video interpolator conditioned on the geometric state. This decomposition aligns with the observation that object trajectories and road geometry evolve on a slower timescale than high-frequency appearance changes such as shadows or turn signal flicker. Consequently, the design of a forecasting system must reflect an understanding that physical coherence operates across multiple temporal scales, and that redundancy in representation can be leveraged to improve efficiency without sacrificing fidelity.

The tension between deterministic geometric modules and stochastic generative modules further defines the architectural frontier. Deterministic 3D reasoning provides guarantees about object permanence and collision-free trajectories but struggles to capture the multimodal nature of human behavior, where a vehicle at an intersection might turn left, proceed straight, or yield. Pure generative models can represent this multimodal distribution but risk hallucinating physically impossible trajectories when unmoored from geometric constraints. A balanced architecture uses a stochastic generative core for behavioral diversity, guided and bounded by a deterministic geometric scaffold that prunes physically infeasible future worlds. This hybrid design philosophy is instantiated in recent approaches that output multiple plausible 3D scene flows, filter them through a physics-based feasibility checker, and then generate corresponding video renderings. The architectural complexity, however, introduces challenges in credit assignment during training, requiring careful loss weighting and curriculum learning strategies that first establish geometric grounding before introducing behavioral multimodality.

3. Physics Coherence and Representation Learning

Physics coherence in the context of autonomous driving scene forecasting extends beyond simple rigid-body dynamics. It encompasses the consistent occlusion relationships among static and dynamic elements, the preservation of object identities within and across frames, the realistic lighting and shadow casting consistent with scene geometry and sun position, and the global kinematic plausibility of multi-agent interactions. A system that generates a future frame in which a pedestrian appears both in front of and behind a parked vehicle from different camera views has violated a fundamental physical law, one that a joint 3D-aligned model can detect and penalize. The operationalization of physics coherence thus requires a suite of differentiable diagnostic measures that are computed alongside the generative loss and serve as self-supervised regularizers during training.

One effective mechanism is to project the generated future frames onto a shared 3D canonical space using estimated depth and camera poses, and then compute consistency losses that measure the discrepancy between multiple observations of the same world point. This approach, borrowed from multi-view stereo and neural rendering literature, forces the generative model to produce frames that correspond to a single, globally consistent 3D scene structure [8]. When combined with optical flow supervision, the system can learn to track points across time and enforce that the 2D motion in the image plane corresponds to a plausible 3D scene flow. This constraint is particularly valuable for disambiguating cases where camera ego-motion and object motion are entangled, a ubiquitous challenge in moving platform perception.

The representation learning agenda for physics-coherent forecasting also benefits from integrating symbolic or semi-symbolic physical knowledge into the neural backbone. Instead of requiring the model to rediscover Newtonian mechanics purely from pixel data, architectural biases such as differentiably simulated moving particles or scene graph nodes with explicit mass, velocity, and extent can be embedded as intermediate representations. The model learns to map raw sensor observations to these structured physical latents and subsequently to generate future frames by propagating the physical state forward in time using learned or hard-coded dynamics. This hybrid neuro-symbolic approach improves sample efficiency and strengthens generalization to novel object configurations, although it risks brittleness when the assumed physical ontology is insufficiently expressive to capture rare interactions such as a tire blowout or a fallen cargo object. The structural trade-off

parallels the broader tension in artificial intelligence between data-driven flexibility and knowledge-driven robustness.

Another important facet of representation learning concerns the integration of map and infrastructure priors. Autonomous vehicles operate within strongly regularized geometric contexts defined by road topology, lane markings, traffic signs, and intersection logic. A forecasting system that jointly generates video and aligns it with a 3D map can use the map as an immutable geometric skeleton that constrains possible future trajectories [7]. The map provides a persistent, survey-grade 3D reference that simplifies the alignment of dynamic objects and reduces the accumulation of drift over long forecast horizons. In scenarios where live map data is unavailable or outdated, the system must dynamically adjust the geometric scaffold, a capability that demands online adaptation loops. This raises further system design questions about the division of responsibility between onboard and off-board mapping infrastructure, as well as the freshness and trustworthiness of prebuilt high-definition maps in rapidly changing environments.

Moreover, physics coherence should be understood as a graded notion rather than a binary criterion. In the real world, certain physical violations are more safety-critical than others. A hallucinated vehicle appearing transiently in a non-drivable region is less dangerous than a failure to predict a child's trajectory crossing the ego-path. Consequently, the loss function and evaluation protocols for physics-coherent forecasting must be risk-aware, assigning higher penalties to geometric misalignment in safety-relevant regions such as the drivable corridor and crosswalks. This risk-weighted learning paradigm connects the model's internal training objectives to the operational safety requirements of the downstream planning and control stack, embedding system-level safety intent directly into the representation.

4. Integration with Autonomous Driving Pipelines

The deployment of a joint generation-alignment forecasting module within a full autonomous vehicle stack introduces real-time operational constraints that fundamentally shape architectural decisions. The forecasting module must consume a multimodal input stream from cameras, lidars, and radars at sensor frame rates, and produce future predictions that interface with motion planning in a latency budget typically under one hundred milliseconds. This demands highly optimized inference pipelines that can leverage hardware accelerators such as tensor processing units and dedicated neural rendering engines. The separation of the 3D alignment component into a scene representation that is updated incrementally, rather than recomputed from scratch for each forecast, emerges as an advantageous design pattern. A persistent 3D world state, maintained via a recurrent neural memory or a transformer-based world model, absorbs new observations as they arrive and provides a consistent geometric substrate for generative queries about the future [4].

The interface between the forecasting module and the downstream planner is a critical integration point that is often underexplored in standalone forecasting benchmarks. A planner that receives only the generated future imagery must still parse that imagery into actionable trajectory constraints, effectively duplicating perception efforts. A more elegant integration exposes the internal 3D aligned representation directly as an occupancy flow or a reachability map that the planner can query. This enables the planner to reason directly in metric world coordinates about the expected future occupancy of road regions, sidestepping the need to interpret generated pixels. The forecasting module thus transitions from a sensor simulator to a world-state predictor, a role that aligns more naturally with safety-critical decision-making. The human-machine interface or remote assistance operator may also benefit from the

generated video as a visual explanation of the vehicle's predicted future, provided the alignment guarantees that what is shown is geometrically faithful.

Calibration of uncertainty presents another integration challenge. The joint system must simultaneously quantify and communicate its uncertainty about both appearance-level details (will the pedestrian wear a red or blue shirt?) and geometric properties (will the pedestrian step onto the road or remain on the sidewalk?). The former is largely inconsequential for planning, while the latter is safety-critical. Effective integration architectures decouple these uncertainty channels, conveying geometric confidence bounds to the planner while retaining full stochasticity in the visual generation for interpretability purposes [16]. Bayesian neural network components or ensemble-based methods can be applied specifically to the geometric head of the model to provide calibrated occupancy uncertainties, even as the video generation head produces samples from a multimodal appearance distribution. This targeted uncertainty decomposition is more computationally efficient and operationally useful than treating the entire model as a monolithic probabilistic black box.

Continuous learning and adaptation mechanisms further complicate integration. The statistical properties of real-world driving environments shift across geographic regions, seasons, and traffic norms. A forecasting model trained predominantly on North American suburban data may exhibit systematically degraded geometric alignment when deployed in dense European urban centers or in monsoon conditions. An integrated system must therefore incorporate online or periodic fine-tuning strategies that adapt the 3D alignment module using self-supervised signals from the vehicle's own sensor stream, without requiring expensive human annotation. Federated learning protocols offer a pathway to share improved world representation capabilities across fleets while maintaining the privacy of individual trip data [19]. The system architecture must be designed with these adaptation hooks from the outset, treating the forecasting module not as a static artifact but as a living system component that evolves with its deployment environment.

5. Robustness, Fairness, and Safety Assurance

The operational safety case for a physics-coherent forecasting system must go beyond average-case accuracy metrics and rigorously interrogate failure modes under distributional shift, adversarial conditions, and demographic disparities. Because the forecasting module directly conditions planning decisions that influence collision avoidance, any systematic bias in prediction quality across different road user categories constitutes a fairness risk. Evidence from pedestrian detection and intention prediction literature shows that recognition models can exhibit disparate performance across demographic groups and mobility device types, raising the concern that a generation-alignment system trained on imbalanced datasets might produce less accurate geometric forecasts for children, wheelchair users, or individuals with non-normative gait patterns [14]. Mitigating such disparities requires deliberate dataset curation, subgroup-specific evaluation protocols, and fairness-aware training objectives that penalize variance in geometric prediction error across protected attributes.

Adversarial robustness is another critical dimension of safety assurance. An attacker who manipulates the input imagery with imperceptible perturbations could potentially cause the forecasting module to hallucinate phantom obstacles or to misalign the predicted 3D scene, leading the planner to execute dangerous evasive maneuvers. The joint alignment constraint offers a potential defense: by enforcing consistency between the generated video and the geometrically inferred 3D structure, the system can detect and reject adversarial inputs that cause physcially implausible latent representations. The explicit geometric head acts as a

verifier that checks for violations of multi-view coherence, providing a layer of invariant reasoning that complements the statistical robustness of the generative backbone. Certifying the forecasting system against adversarial threats will likely require a combination of formal verification of the geometric consistency module and empirical stress-testing via physics-aware adversarial generation.

The safety assurance case also demands rigorous corner-case evaluation. The joint system must be tested on scenarios where the physical world undergoes rapid, atypical changes, such as a sudden road collapse, a traffic light malfunction causing conflicting greens, or an animal darting from an unobserved region. In such cases, a purely generative model may extrapolate smoothly from training data and fail to represent the physical discontinuity, whereas the 3D alignment module, relying on sensor observations, might correctly flag an inconsistency with the prior map. The system architecture should be designed to prioritize sensor-anchored geometric evidence over generative priors when a significant discrepancy is detected, effectively entering a conservative fallback mode that passes high uncertainty to the planner. Operational design domain (ODD) expansion therefore requires not just more training data but also a system-level meta-cognitive capability to recognize when the world has deviated from the model's training manifold.

From a broader socio-technical perspective, the deployment of generative forecasting systems intersects with public trust, liability, and regulatory approval. A forecasting system that is perceived as a black box oracle, no matter how physically coherent internally, may face public resistance if involved in high-profile incidents. The joint architecture provides a transparency mechanism: the generated video can be presented as a human-interpretable representation of the vehicle's predicted future, while the 3D occupancy and flow outputs can be scrutinized by independent auditing tools. Establishing standardized benchmarks and third-party evaluation frameworks for physics-coherent forecasting is an essential step toward regulatory acceptance. These benchmarks must move beyond simplistic image quality metrics such as FID or PSNR and incorporate geometric alignment scores, collision rate estimates, and fairness disparity indices, reflecting the multidimensional safety requirements of real-world deployment.

6. Governance, Policy, and Sustainability

The governance of generative forecasting technologies for autonomous driving is situated at the intersection of artificial intelligence regulation, transportation safety law, and environmental policy. The introduction of large-scale image-to-video models into safety-critical vehicular systems triggers scrutiny under emerging AI accountability frameworks, such as the European Union's AI Act, which classifies certain autonomous vehicle functions as high-risk applications requiring conformity assessments, transparency documentation, and human oversight provisions. The joint generation-alignment architecture, by making geometric reasoning explicit and auditable, may facilitate compliance with transparency requirements more readily than opaque end-to-end black-box predictors. Demonstrating that generated video forecasts are geometrically faithful to an independently verifiable 3D world model could become a key element of the technical file submitted to regulatory bodies. Policy development should therefore encourage the publication of alignment and consistency audit trails alongside model outputs.

Issues of data governance and privacy also feature prominently because training generative video models at the scale required for diverse driving environments involves ingesting vast quantities of street-level imagery that may capture personally identifiable information of

pedestrians, cyclists, and vehicle occupants. The 3D alignment component, while primarily concerned with geometric structure, relies on feature representations extracted from this same imagery, raising questions about the extent to which privacy-preserving techniques such as differential privacy or federated learning can be integrated without degrading alignment accuracy [19]. Synthetic data generation, enabled by the very forecasting technology under discussion, creates a circular dynamic: forecasting models can generate privacy-compliant training data for other perception models, but this synthetic data must itself be rigorously validated for physical coherence and statistical representativeness. Governance frameworks should establish guidelines for the acceptable use of recycled synthetic data in safety-critical training pipelines to avoid data quality degradation over successive generations.

The environmental sustainability of training and deploying large-scale generative forecasting models represents an additional policy concern that is frequently overshadowed by safety discussions. The computational requirements of diffusion-based video generators and neural radiance field renderings are substantial, and scaling these models to accommodate the vast diversity of global driving conditions would result in significant energy consumption and associated carbon emissions [18]. The architectural choice to jointly optimize generation and 3D alignment may yield net energy savings if the geometric inductive biases reduce the amount of data and training iterations needed to achieve physics-coherent behavior. However, this efficiency hypothesis must be empirically validated through comprehensive lifecycle assessments that account for both training and inference energy costs. Future policy instruments, such as carbon footprint disclosure mandates for AI models used in public infrastructure, could incentivize the development of more efficient architectures and encourage the sharing of pretrained world models across manufacturers to amortize training costs.

Intellectual property and liability attribution also require clarity. When a jointly generated and aligned forecast contributes to a collision, the question of whether the generative component, the geometric alignment module, or the planner is the proximal cause becomes non-trivial. The system architecture influences this attribution: a decoupled design allows fault isolation and component-level certification, whereas a tightly integrated design may diffuse responsibility across a single monolithic model. From a governance standpoint, requiring modular architectures with well-defined interfaces for safety-critical functions, including forecasting, could simplify certification and post-incident analysis. Furthermore, the training data provenance and any embedded biases inherited from proprietary datasets may become subjects of legal discovery. Transparent documentation of training data composition, geometric alignment error characteristics, and fairness audit results will likely become standard practice, paralleling the safety case documentation already required for traditional automotive components.

7. Conclusion

The pursuit of physics-coherent autonomous driving scene forecasting through joint image-to-video generation and 3D world alignment represents a significant architectural convergence of generative modeling, geometric computer vision, and safety-critical systems engineering. This paper has argued that moving beyond disjointed prediction pipelines toward unified frameworks that jointly reason about appearance and geometry can address fundamental physical consistency failures that currently limit the trustworthiness of autonomous forecasting. The system-level analysis presented here underscores that the design of such frameworks involves navigating a rich space of trade-offs between generative expressiveness

and deterministic reasoning, between computational feasibility and physical fidelity, and between centralized training efficiency and on-device adaptation capability. The architectural patterns that embed differentiable 3D alignment within video generation backbones, supported by risk-aware loss formulations and uncertainty decomposition, offer a technically viable path forward, though they are not without their own failure modes and computational costs.

Crucially, the success of these systems cannot be evaluated solely through narrow technical benchmarks. Their deployment into real-world transportation ecosystems implicates a broader set of governance, fairness, and sustainability requirements that must be addressed concurrently with algorithmic innovation. The need for subgroup-fair geometric prediction, adversarial robustness, transparent auditability, and environmentally responsible training is not peripheral to the research agenda but constitutive of it. The joint generation-alignment paradigm, if developed with these socio-technical dimensions in mind, can provide both a more physically grounded forecasting capability and a more accountable system architecture. Future research should prioritize longitudinal field studies that assess the reliability of these models across diverse operational design domains, the development of standardized physics-coherence certification benchmarks, and the exploration of compact and energy-efficient architectural variants. By integrating systems thinking with algorithmic depth, the research community can steer autonomous forecasting toward a future that is not only visually convincing but also geometrically honest, physically safe, and socially responsible.

References

1. Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., ... & De Souza, A. F. (2021). Self-driving cars: A survey. Expert Systems with Applications, 165, 113816.

2. Finn, C., Goodfellow, I., & Levine, S. (2016). Unsupervised learning for physical interaction through video prediction. In Advances in neural information processing systems (pp. 64-72).

3. Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision (pp. 405-421). Springer.

4. Hu, A., Natan, A., Salvador, J., Ramanan, D., & Gutfreund, D. (2023). GAIA-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080.

5. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., ... & Salimans, T. (2022). Imagen Video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303.

6. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., ... & Dai, J. (2022). BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In European conference on computer vision (pp. 1-18). Springer.

7. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11621-11631).

8. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707.

9. Salzmann, T., Ivanovic, B., Chakravarty, P., & Pavone, M. (2020). Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In European Conference on Computer Vision (pp. 683-700). Springer.

10. Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In Advances in neural information processing systems (pp. 613-621).

11. Xiong, Z., Song, Y., He, L., Xiong, W., Yuan, Y., Qiao, F., & Jacobs, N. (2026). PhysAlign: Physics-Coherent Image-to-Video Generation through Feature and 3D Representation Alignment. arXiv preprint arXiv:2603.13770.

12. Seo, H., Kim, S., Shin, J., & Kim, J. (2023). Dense depth-guided generalizable NeRF. IEEE Robotics and Automation Letters, 8(5), 2896-2903.

13. Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.

14. Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. arXiv preprint arXiv:1902.11097.

15. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 23-30). IEEE.

16. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in neural information processing systems (pp. 5574-5584).

17. Ryan, M. (2020). The future of transportation: ethical, legal, social and economic impacts of self-driving vehicles in the year 2025. Science and Engineering Ethics, 26, 1185-1208.

18. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th annual meeting of the Association for Computational Linguistics (pp. 3645-3650).

19. 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.

20. Prakash, A., Chitta, K., & Geiger, A. (2021). Multi-modal fusion transformer for end-to-end autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7077-7087).