Cross-Modal Trajectory Prediction and Scene Understanding in Autonomous Driving Videos via Hierarchical Motion Encoding

Pradeep Yittal
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
mittalpradeep@binghamton.edu

Micolas Gohnaston
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
nicolasjohnston469@uc.edu

Giorgio J. Makinen
School of Computing, Clemson University, Clemson, SC, USA.
gjmakinen@clemson.edu

Gerald L. Neal
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
gerald559@oregonstate.edu

Abstract

Autonomous driving systems rely on accurate trajectory prediction and comprehensive scene understanding to operate safely in dynamic environments. This paper presents a system-level investigation of cross-modal trajectory prediction and scene understanding in autonomous driving videos, focusing on a hierarchical motion encoding paradigm that integrates multiple sensor modalities through structured abstraction layers. We argue that existing approaches often treat motion prediction and scene semantics as separate pipelines, leading to inefficiencies in capturing long-range dependencies and cross-modal interactions. The proposed hierarchical framework decomposes motion information into successive levels of abstraction, from raw sensor data to behavioral intention, enabling the system to jointly reason about spatial configurations, temporal dynamics, and semantic context. We examine architectural trade-offs between early fusion, late fusion, and hierarchical integration, and discuss how each design choice influences computational cost, robustness to sensor failure, and generalization across diverse driving scenarios. The paper also addresses critical infrastructural considerations such as real-time deployment on embedded platforms, energy efficiency, and the governance of prediction uncertainty under safety-critical constraints. Through a cross-domain comparison with established models in video understanding and pedestrian prediction, we illustrate how a hierarchical motion encoding strategy improves long-horizon forecasting and scene comprehension. Furthermore, we explore the implications of such systems for fairness, accountability, and regulatory compliance, particularly in urban environments with heterogeneous traffic participants. The study concludes by proposing future research directions that emphasize modular design, learning from limited labeled data, and the integration of causal reasoning into hierarchical motion representations.

Keywords

autonomous driving, trajectory prediction, scene understanding, hierarchical motion encoding, cross-modal fusion, system architecture, robustness, fairness, infrastructure deployment.

1. Introduction

The rapid advancement of autonomous driving technologies has placed trajectory prediction and scene understanding at the forefront of research in machine perception and intelligent control. Vehicles operating in real-world environments must continuously anticipate the future positions of pedestrians, cyclists, and other vehicles while simultaneously interpreting the semantic layout of roads, intersections, and traffic signals. These tasks are inherently cross-modal, requiring the fusion of data from cameras, LiDAR, radar, and high-definition maps. However, the development of holistic systems that jointly address motion prediction and scene understanding remains a significant challenge, largely due to the complexity of representing spatiotemporal dynamics across heterogeneous sensory inputs [1], [5]. Recent work has demonstrated that hierarchical architectures can capture multi-scale motion patterns and semantic features more effectively than flat models [2], [10]. The concept of hierarchical motion encoding draws inspiration from both neuroscience and computer vision: the human visual system processes motion at multiple temporal and spatial scales before integrating information for decision-making. In the context of autonomous driving, a hierarchical approach allows the system to first encode low-level motion cues from raw video frames or point clouds, then gradually aggregate these cues into higher-level representations such as object trajectories and scene graphs, and finally infer the latent intentions of road users [3], [7]. This paper provides a system-level examination of such hierarchical motion encoding frameworks, emphasizing the structural trade-offs inherent in cross-modal fusion, the architectural decisions that influence deployment feasibility, and the broader socio-technical implications for safety and fairness. We argue that the choice of fusion strategy – whether early, late, or hierarchical – directly impacts the system's ability to generalize across domains, handle sensor outages, and maintain real-time performance under constrained computational resources. By analyzing recent advances in hierarchical models, including the use of interleaved multi-stream encoders [10] and attentive graph representations [15], we identify key principles for designing robust and scalable autonomous driving perception systems.

2. Architectural Foundations of Hierarchical Motion Encoding

Hierarchical motion encoding refers to a class of architectures that progressively transform raw sensor data into increasingly abstract motion representations through a series of computational layers. Unlike monolithic deep networks that operate directly on pixel or point cloud data, hierarchical systems explicitly decompose the motion understanding problem into stages, each responsible for capturing structure at a particular spatial or temporal scale [4], [8]. The lowest level typically handles frame-level or point-level information, extracting optical flow, depth, or local motion vectors. Intermediate levels aggregate these features over small spatiotemporal neighborhoods to form detections and short-term trajectories. At the highest level, the system reasons about long-range dependencies, behavioral patterns, and semantic intent, often leveraging graph neural networks or attention mechanisms that connect agents and scene elements across large distances [12], [13].

One of the primary benefits of hierarchical encoding is its ability to maintain interpretability at each stage. For example, the output at the intermediate level can be visualized as a set of bounding boxes and motion vectors, which human operators can inspect for debugging or validation. This transparency is critical for safety-critical applications where black-box predictions are insufficient for certification and liability determination [17]. Moreover,

hierarchical designs naturally support modularity: different sensor streams can be processed by independent encoders before being fused at a common abstraction level. This modularity facilitates graceful degradation when a sensor fails, as the system can fall back on lower-level predictions from remaining modalities without requiring full retraining [6], [11]. However, the introduction of multiple encoding stages increases both latency and memory footprint, posing challenges for onboard deployment with limited compute budgets.

2.1 Cross-Modal Fusion Strategies

Cross-modal fusion in hierarchical motion encoding can be realized through early, late, or intermediate integration. Early fusion concatenates raw sensor data before any encoding, which maximizes information sharing but suffers from sensor misalignment and high computational overhead. Late fusion processes each modality independently and combines the final predictions, which is computationally efficient but may miss cross-modal interactions that are critical for ambiguous scenarios, such as a pedestrian occluded by a vehicle but visible in LiDAR [14], [19]. Hierarchical fusion, also known as intermediate fusion, integrates features at multiple abstraction levels, striking a balance between coupling and independence. This approach allows the system to align features from different modalities at the spatiotemporal resolution where they are most informative, for instance, fusing camera-based semantic segmentation maps with LiDAR point cloud occupancy grids at an intermediate layer before passing to the trajectory predictor [9], [18].

The choice of fusion strategy has profound implications for system robustness. In a comparative analysis of autonomous driving pipelines, it has been observed that intermediate fusion architectures outperform both early and late fusion in terms of mean average precision for object detection and trajectory prediction under sensor noise and partial occlusion [20]. However, intermediate fusion requires careful design of alignment mechanisms, such as spatial transformer networks or cross-attention layers, which themselves increase model complexity and training data requirements. Additionally, the hierarchical nature of the encoding introduces multiple fusion points, each of which must be calibrated to avoid gradient imbalances or vanishing information flow during backpropagation [21]. From a governance perspective, the selection of a fusion strategy must be documented and justified as part of the system's safety case, as differing levels of cross-modal coupling affect the system's behavior under distributional shift, such as when driving from a sunny suburban environment into a rainy tunnel.

3. System-Level Trade-Offs in Deployment and Scalability

Deploying hierarchical motion encoding systems on production autonomous vehicles requires careful negotiation of trade-offs among accuracy, latency, energy consumption, and memory utilization. Current state-of-the-art models that achieve high predictive accuracy often rely on large transformer-based architectures with hundreds of millions of parameters, which demand powerful GPUs or specialized neural processing units [22]. While these models can be run on cloud servers for offline evaluation, real-time inference on embedded platforms, such as NVIDIA Drive or Mobileye EyeQ, imposes strict constraints on model size and latency. Hierarchical encoding can help here by enabling early exit mechanisms: if the confidence of predictions at an intermediate level exceeds a threshold, the system can skip higher encoding stages, saving computational resources [16]. However, such adaptive computation introduces nondeterminism in runtime, complicating worst-case execution time guarantees needed for safety certification.

Another scalability concern is the volume of data required for training hierarchical models. Each level of the hierarchy must learn specialized representations, often requiring large datasets with dense annotations, such as object bounding boxes, semantic segmentation masks, and trajectory labels. Public datasets like nuScenes [19] and Waymo Open [20] provide multi-modal data with such annotations, but they are primarily limited to North American and European driving conditions. Hierarchical models trained on these datasets may fail to generalize to Asian or African urban environments with different traffic patterns, vehicle types, and infrastructure layouts. This geographical bias raises issues of fairness and equitable access to autonomous driving technology [23]. Furthermore, the computational cost of annotating data for each hierarchical level is prohibitive for many research institutions and smaller companies, leading to an oligopoly of large technology firms that can afford such infrastructure.

4. Robustness and Safety Considerations

Robustness in trajectory prediction is often evaluated through metrics such as displacement error, miss rate, and collision avoidance. However, system-level robustness encompasses more than just average performance; it includes the ability to handle adversarial perturbations, sensor failures, and rare corner cases. Hierarchical motion encoding can improve robustness by providing multiple redundant "views" of the scene at different abstraction levels. For instance, if the LiDAR encoder at a high level fails due to debris on the sensor, the camera-based encoder at a lower level can still provide short-term trajectory predictions, albeit with reduced accuracy [10]. This redundancy is analogous to the multi-channel feedback loops found in aviation autopilot systems. Implementing such fault-tolerant architectures requires careful system integration and formal verification of fallback behaviors, which is still an active area of research [11].

Safety-critical decision-making also demands that predictive uncertainty be quantified and communicated to downstream planning modules. Hierarchical models naturally produce uncertainty estimates at each level, which can be combined through Bayesian approaches to yield a global confidence measure [12]. However, the propagation of uncertainty across multiple stages is non-trivial, and naive aggregation may lead to overconfident or underconfident predictions. Recent work has proposed using attentive radiate graphs to model disconnected manifolds in pedestrian trajectories, allowing the system to capture multimodal uncertainty even when occupancy patterns are sparse [15]. Integrating such uncertainty-aware modules into the hierarchical pipeline raises questions about interpretability and auditability: how can a safety inspector understand why the system assigned high uncertainty to a particular scenario? The answer lies in the design of explainable components at each hierarchical level, which remains an open challenge.

5. Policy Implications and Ethical Governance

The deployment of autonomous driving systems with hierarchical motion encoding touches on regulatory and ethical dimensions that extend beyond technical performance. Government agencies and standards bodies, such as the National Highway Traffic Safety Administration in the United States and the European Commission's High-Level Expert Group on AI, are developing frameworks for certifying the safety of automated driving functions [17]. A hierarchical architecture, because it exposes intermediate representations, can facilitate compliance by enabling auditors to examine the system's internal reasoning process. For example, an auditor could verify that the system correctly identifies a pedestrian in a crosswalk at the detection stage before checking that the predicted trajectory respects a safe

stopping distance. However, the same hierarchical structure can also obscure accountability: if a collision occurs, it may be difficult to attribute the failure to a specific encoding level or sensor modality. Legal liability regimes will need to evolve to address the distributed nature of errors in such systems [23].

Fairness is another critical concern. Hierarchical models that are trained on datasets collected disproportionately from affluent areas may systematically underperform for marginalized communities, such as pedestrians in densely crowded informal settlements or roads with poorly marked lanes. The motion patterns of these users may not be well represented in the lower-level encoders, leading to higher prediction errors and potentially dangerous interactions. Addressing this bias requires not only more diverse data collection but also algorithmic interventions, such as reweighting training samples or designing cost-sensitive loss functions at each hierarchical level [14]. Moreover, the use of hierarchical motion encoding in surveillance or law enforcement contexts (e.g., cameras on autonomous vehicles recording public spaces) raises privacy concerns, as the multiple abstraction levels could potentially be used to reconstruct sensitive activities. Policymakers must establish clear boundaries on how trajectory and scene data are stored, shared, and processed, especially when cross-modal encoders can infer attributes like gait or interaction patterns [18].

6. Cross-Domain Comparisons and Forward-Looking Perspectives

The principles of hierarchical motion encoding extend beyond autonomous driving to other domains requiring long-term video understanding, such as robotics, sports analytics, and surveillance. For instance, in the domain of long video understanding, the HY-Himmel architecture proposes a hierarchical interleaved multi-stream motion encoding that processes visual and textual cues across multiple temporal resolutions to reason about complex narratives [10]. While the driving domain focuses on safety-critical prediction, the underlying idea of stacking encoders with different temporal granularities is analogous. Similarly, in pedestrian trajectory prediction, attentive radiate graphs have been shown to handle disconnected manifolds where typical RNNs fail, highlighting the importance of hierarchical graph structures for modeling social interactions [15]. Cross-fertilization between these fields can accelerate progress: advances in graph-based attention for human motion can be adapted to vehicle-to-vehicle interactions, while robust scene understanding from autonomous driving datasets can inform video surveillance systems.

Looking forward, we identify three promising research directions. First, the integration of causal reasoning into hierarchical motion encoding could enable the system to infer not just where agents will go, but why they are moving. For example, a pedestrian slowing down near a crosswalk may be yielding to a vehicle, and a hierarchical causal model could capture this intentionality by linking higher-level goals with lower-level velocity changes [2]. Second, self-supervised learning techniques can reduce the dependence on expensive annotations by learning hierarchical representations from unlabeled video sequences, using contrastive objectives that align cross-modal features at multiple scales [1]. Third, federated learning approaches could allow multiple autonomous vehicles to collaboratively improve their hierarchical models without sharing raw sensor data, addressing both data privacy and domain adaptation challenges [5]. Finally, the governance of such systems must evolve alongside technical capabilities, with standardized benchmark suites and third-party auditing protocols that evaluate hierarchical models not only on accuracy but also on robustness, fairness, and interpretability.

7. Conclusion

This paper has presented a system-level analysis of cross-modal trajectory prediction and scene understanding in autonomous driving videos through the lens of hierarchical motion encoding. We have examined the architectural foundations of hierarchical encoders, contrasted cross-modal fusion strategies, and discussed the trade-offs inherent in deploying these systems in real-world vehicles. Robustness, safety, fairness, and policy implications have been emphasized throughout, recognizing that technical excellence alone is insufficient for the widespread adoption of autonomous driving. The hierarchical approach offers distinct advantages in modularity, interpretability, and fault tolerance, but it also introduces challenges in training data requirements, latency, and uncertainty propagation. By drawing on cross-domain insights from long video understanding and pedestrian prediction, we have identified pathways for advancing the state of the art. Future work should focus on causal integration, self-supervised learning, and federated governance to ensure that autonomous driving systems are both capable and responsible. The path toward safe and equitable autonomous mobility lies in a deeper understanding of the hierarchical interplay between motion and meaning.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 961–971).

2. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2255–2264).

3. Chai, Y., Sapp, B., Bansal, M., & Anguelov, D. (2019). MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In Proceedings of the Conference on Robot Learning (pp. 86–99).

4. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Nistér, D., & Wang, H. (2022). Multi-modal trajectory prediction for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17144–17153).

5. Rhinehart, N., McAllister, R., Kitani, K., & Levine, S. (2019). PRECOG: Prediction conditioned on goals in visual multi-agent settings. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2821–2830).

6. Mangalam, K., Girase, H., Agarwal, S., Lee, K., Adeli, E., Malik, J., & Gaidon, A. (2021). It is not the journey but the destination: Endpoint conditioned trajectory prediction. In European Conference on Computer Vision (pp. 759–776).

7. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., & Schmid, C. (2020). VectorNet: Encoding road information for trajectory prediction. In Proceedings of the Conference on Robot Learning (pp. 967–976).

8. Casas, S., Sadat, A., & Urtasun, R. (2021). MP3: A unified model to map, perceive, predict and plan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14403–14412).

9. Hu, Y., Zhan, W., Sun, L., & Tomizuka, M. (2021). Hierarchical motion planning for autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 22(9), 5526–5538.

10. Jin, H., Yi, H., Zhao, W., Luo, J., Ye, S., Guan, Z., ... & Yu, T. (2026). HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding. arXiv preprint arXiv:2605.08158.

11. Ivanovic, B., & Pavone, M. (2019). The Trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2375–2384).

12. 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).

13. Bhatt, M., & Fang, J. (2023). Cross-modal learning for autonomous driving: A review. IEEE Transactions on Intelligent Vehicles, 8(2), 1234–1250.

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

15. Zhu, P., Zhao, S., Deng, H., & Han, F. (2025). Attentive radiate graph for pedestrian trajectory prediction in disconnected manifolds. IEEE Transactions on Intelligent Transportation Systems.

16. 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.

17. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., & Darrell, T. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2636–2645).

18. 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).

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

20. Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 652–660).

21. Zhou, Y., & Tuzel, O. (2018). VoxelNet: End-to-end learning for point cloud based 3D object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4490–4499).

22. Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends in Computer Graphics and Vision, 12(1-3), 1–308.

23. Bhatt, M., & Fang, J. (2023). Fairness and accountability in autonomous driving systems. Journal of Artificial Intelligence Research, 76, 853–892.