Modification and Responses to the Reviewers' Comments and Suggestions
Manuscript ESWA-D-25-28448

Learning and Predicting Traffic Conflicts in Mixed Traffic: A Spatiotemporal Graph Neural Network with Manifold Similarity Learning

Zongshi Liu, Guojian Zou, Ting Wang, Meiting Tu, Hongwei Wang, Ye Li

AE: There are conflicting review reports, even the recommendation to reject the paper. Against this backdrop it is crucial to carefully address all the comments, especially the critical ones.

Dear Professor Eklund,
We are very grateful to the Editor and the Reviewers for spending precious time reviewing our work and providing useful comments. We have carefully considered and responded to each of the comments from the reviewers. Below we briefly summarize the major revisions that have been made to improve the quality and clarity of the manuscript:

- Clarified the methodological novelty and positioning of MS-STGNet relative to existing manifold-based traffic models and STGAT-type adaptive graph networks, including expanded discussions in Sections 1, 2.2, 2.3, 5.3.1, and 6.5.
- Strengthened the justification of our simulation-based framework and data choices, explaining the limitations of current mixed CAV–HDV datasets and detailing our "real-trajectory calibration + large-scale simulation" strategy, together with an explicit statement of validation limitations and future real-world testing plans in the Conclusion.
- Enhanced the analysis of model performance, stability, and sensitivity by reporting mean ± standard deviation over five independent runs (with statistical significance at the 5% level), clarifying the posterior-probability analyses (Section 6.8), and adding new sensitivity-style results on traffic volume, CAV penetration, and speed separation in Section 6.9 and Appendix C.
- Added a computational-cost analysis (Section 6.6) comparing GPU memory usage and parameter counts across all deep learning baselines and MS-STGNet, thereby demonstrating that our model achieves superior performance with a computational burden comparable to or lower than STGCN and STGAT.
- Provided a clearer discussion of influential traffic features for conflict prediction (e.g., CAV penetration, traffic volume, speed dispersion patterns, and vehicle composition), and how MS-STGNet aligns its predicted risk with these structures.
- Streamlined and updated the literature review, reorganized some technical details into appendices, refined mathematical notation, and polished the language throughout to improve readability and consistency.
- Expanded the discussion of limitations and future research directions, especially regarding scenario generalizability, online manifold updating, graded conflict severity modelling, and large-scale real-world deployment.

Explanations of what we have changed in response to the reviewers' concerns are given point by point in the following pages. The changes in the revised manuscript have highlighted in blue. We hope these changes will strengthen our manuscript.

Comments from Reviewer #1:

Comment 1:

Firstly, the novelty of the proposed framework is somewhat limited. Although the integration of manifold similarity into a spatiotemporal graph neural network is an interesting idea, the concept of using manifold learning for traffic state modeling is not entirely new and has been explored in previous studies. The manuscript needs to provide a more detailed comparison with existing methods to clearly highlight the unique contributions of the proposed MS-STGNet framework.

Response to Comment 1:

We sincerely thank the reviewer for these thoughtful comments on the novelty of MS-STGNet and its relation to existing manifold-based traffic models and STGAT-type adaptive graph networks.

In the revised manuscript, we have clarified our contributions at both the problem and method levels. From the problem perspective, we emphasize that our primary goal is real-time conflict prediction in mixed CAV–HDV freeway traffic, a setting where existing work is still limited. To ensure robustness in this new safety-critical application, we deliberately build on mature components (residual CNN, TCN, spatiotemporal GNNs) while introducing a manifold-similarity graph as a physically meaningful prior rather than proposing an entirely new architecture for its own sake.

From the methodological perspective, we now explicitly distinguish our approach from prior manifold-learning studies and STGAT-type adaptive adjacency mechanisms. Section 2.3 has been expanded to include recent manifold-based traffic-flow and safety studies and to clarify that these works mainly use manifold embeddings for clustering, visualization, or as features in conventional models, without embedding manifold-based traffic-state similarity into a spatiotemporal GNN for online conflict prediction. In contrast, MS-STGNet integrates a pre-computed manifold similarity matrix, derived from historical traffic states, as an interpretable prior that constrains adaptive adjacency learning for mixed CAV–HDV conflicts.

We also revise Section 2.2 and Section 5.3.1 to clearly contrast our manifold-similarity graph with standard STGAT mechanisms: instead of learning adjacency solely from instantaneous node features at each time step, MS-STGNet initializes the graph from manifold distances computed offline and then performs lightweight adaptive refinement. This design ties the learned graph to physically meaningful traffic-state geometry while keeping the per-iteration computational cost comparable to standard STGNNs. Finally, in Section 6.5 we explicitly highlight that the manifold-similarity prior contributes to the reduction of false alarm rates and improved robustness compared with STGCN and STGAT, especially at medium-to-high CAV penetration rates.

We hope these revisions and clarifications make the unique contributions and methodological positioning of MS-STGNet more evident. Below is the revised version:

Revised:

In Section 1 (Page 2 Line 45—49, Page 3 Line 10—14):

Second, we propose MS-STGNet, a spatiotemporal graph neural network that fuses

physical adjacency and semantic features for traffic conflict prediction in mixed CAV–HDV traffic. The framework intentionally builds on mature components (e.g., residual CNN and TCN) to ensure robustness in this new application setting, while introducing a manifold-similarity graph as a physically meaningful prior for adaptive adjacency, which has not been explored in existing mixed-traffic conflict prediction models.

In MS-STGNet, a manifold similarity graph module has been developed and implemented. By leveraging a similarity matrix derived from traffic state data within the manifold space, we provide prior knowledge regarding the evolution of traffic states. The manifold-similarity module incorporates a broader array of traffic-flow attributes during neighbor selection and uses a pre-computed manifold similarity matrix as an interpretable structural prior, thereby reducing the propensity for false-positive conflict-event predictions.

In Section 2.2 (Page 4 Line 41—44):

In addition, existing spatiotemporal graph-based safety models typically define spatial dependencies through fixed adjacency matrices or adaptive attention mechanisms in the original feature space, and rarely exploit manifold-based traffic-state similarity as an explicit prior, particularly in mixed CAV–HDV traffic environments.

In Section 2.3 (Page 5 Line 26—29):

Additionally, few studies have attempted to integrate the concept of state transitions in manifold learning into deep learning frameworks, and, to the best of our knowledge, none has embedded manifold-based traffic-state similarity into a spatiotemporal graph neural network for real-time conflict prediction in mixed CAV–HDV traffic.

In Section 5.3.1 (Page 12 Line 1—10):

Conceptually, the proposed manifold-similarity graph plays a role that is related to, but distinct from, the adaptive adjacency mechanisms used in STGAT-type models. In conventional STGAT, edge weights are learned solely from instantaneous node features via attention, and the adjacency matrix is dynamically reconstructed at each time step. In MS-STGNet, the adjacency structure is instead initialized from manifold distances computed over historical traffic states, which encode long-term traffic-flow evolution and physically meaningful similarity between spatiotemporal patterns. The subsequent adaptive update in MSGNet refines this manifold-based prior rather than discarding it. This separation between a manifold-informed prior graph that reflects the geometric structure of traffic dynamics and a lightweight adaptive refinement brings two benefits: it constrains the learned graph to remain consistent with empirical traffic-state geometry, and it limits the additional per-iteration cost compared with fully attention-based dynamic graphs, keeping the overall complexity comparable to that of standard STGNN models.

In Section 6.5 (Page 18 Line 31—43):

These empirical results also clarify how MS-STGNet differs in practice from STGAT-type adaptive graph models. Although both approaches employ graph-based representations, STGAT relies on feature-driven attention to construct adjacency at each time step, which can be sensitive to local fluctuations in highly imbalanced conflict datasets. By contrast, MS-

STGNet constrains the adaptive graph updates within a manifold-similarity prior derived from historical traffic states. As the market penetration of CAVs increases and pronounced speed separation emerges, this manifold-informed prior helps the model avoid spuriously high conflict probabilities in non-conflict regions, leading to consistently lower false alarm rates and more stable performance across all penetration scenarios. In this sense, our findings are consistent with previous studies showing that graph-based spatiotemporal models such as STGCN and STGAT outperform traditional machine-learning and sequence models in traffic prediction tasks, while further extending them by explicitly incorporating a manifold-based state similarity prior into the adaptive graph learning process. At the same time, our results complement recent manifold-learning approaches for traffic state analysis by demonstrating that manifold-informed similarity can be embedded into deep spatiotemporal graph networks to improve conflict prediction in mixed CAV–HDV freeway traffic.

Comment 2:

Secondly, the experimental setup and validation process lack sufficient rigor. The simulation environment, while realistic, is based on predefined parameters and assumptions that may not fully capture the complexities of real-world mixed traffic conditions. The manuscript should include more comprehensive validation using real-world traffic data to demonstrate the practical applicability and robustness of the proposed model.

Response to Comment 2:

We sincerely appreciate this important comment. At the early stage of this study, our initial intention was indeed to develop and validate MS-STGNet directly on real-world mixed CAV–HDV data. However, after a thorough review of existing datasets, we found that currently available data sources cannot simultaneously meet the two core requirements of our problem: 1) truly mixed CAV–HDV traffic, and 2) long, spatially continuous freeway segments with macroscopic measurements (flow, speed, occupancy) suitable for segment-level conflict prediction over continuous time series.

On the one hand, classical trajectory datasets such as NGSIM and highD contain only human-driven vehicles and therefore do not match our target mixed-traffic scenario. Nevertheless, to ensure that our simulation is not based on purely theoretical assumptions, we calibrated the human-driven car-following model directly on highD freeway trajectories, so that HDV behaviour in the simulation reflects empirically observed acceleration, deceleration, and headway patterns rather than arbitrary parameter choices.

On the other hand, recent autonomous-vehicle datasets such as the Lyft Level 5 AV Dataset (Houston et al., 2021), nuScenes (Caesar et al., 2020), and the Waymo Open Dataset (Sun et al., 2020) do provide mixed traffic with AVs/CAVs, but their structure is not well suited to our macroscopic conflict-prediction task. As summarized in Table 1 of Hu et al. (2022), these AV datasets are organized into short trajectory segments: Waymo comprises 1,000 segments with a temporal resolution of 0.1 s and a typical segment length of 20 s; Lyft Level 5 contains 366 segments at 0.2 s resolution and 25–45 s duration; and nuScenes includes 1,000 segments with 0.5 s resolution and 20 s duration. These segments are collected from the viewpoint of individual AVs and are neither spatially contiguous along a single freeway facility nor

temporally continuous over long periods. Hu et al. (2022) explicitly note that substantial preprocessing and reconstruction are required even to obtain usable car-following trajectories from these segment-based recordings, and that the resulting data remain fragmented in space and time for macroscopic analyses.

Table 1. Overview of three AV trajectory dataset.

Dataset	Number of segments	Resolution (s)	Length of each segment (s)
Waymo	1000	0.1	20
Lyft	366	0.2	25–45
nuScenes	1000	0.5	20

[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).

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., ... & Ondruska, P. (2021, October). One thousand and one hours: Self-driving motion prediction dataset. In Conference on Robot Learning (pp. 409-418). PMLR.

Hu, X., Zheng, Z., Chen, D., Zhang, X., & Sun, J. (2022). Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research. *Transportation Research Part C: Emerging Technologies*, *134*, 103490.

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).]

Similarly, Zhang et al. (2025) show that such AV trajectory datasets are particularly suitable for microscopic behaviour analysis and data-driven stochastic fundamental-diagram modelling, but they are inherently sparse in space and time and therefore not directly aligned with macroscopic, segment-level modelling of traffic states over extended freeway sections. In our study, however, the prediction target is whether a given freeway segment will experience a conflict within a continuous time series, based on macroscopic indicators (flow, speed, occupancy) monitored along a 14 km stretch. This requires long, contiguous observations along one facility, which current AV datasets do not provide.

[Zhang, X., Yang, K., Sun, J., & Sun, J. (2025). Stochastic fundamental diagram modeling of mixed traffic flow: A data-driven approach. *Transportation Research Part C: Emerging Technologies*, *179*, 105279.]

These limitations are consistent with the broader challenges identified in recent reviews of machine-learning-based crash prediction. For example, Ali et al. (2024) point out that empirical safety studies for mixed CAV–HDV environments are still scarce, that most real-time crash and conflict prediction models are developed for conventional freeway or urban networks without CAVs, and that many studies necessarily rely on simulation or indirectly inferred data due to the lack of suitable field datasets. Against this background, we adopted a hybrid strategy that combines empirical calibration with large-scale simulation.

[Ali, Y., Hussain, F., & Haque, M. M. (2024). Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention*, *194*, 107378.]

Given these data limitations, we adopted a hybrid “real-trajectory calibration + simulation-

based testing" strategy. At the microscopic level, we construct mixed CAV–HDV traffic with multiple penetration rates (10%, 30%, 50%, 70%, and 90%) and generate trajectories at 0.2s resolution. Using these trajectories, we compute widely used surrogate safety measures TTC, DRAC, and DDR—to identify both longitudinal and lateral conflicts. At the macroscopic level, we aggregate the simulated data along a 14 km four-lane freeway with ramps and extract continuous time series of flow, speed, and occupancy for each segment, paired with the conflict/non-conflict labels derived from TTC/DRAC/DDR. This design allows us to study conflict prediction as a segment-level time-series classification problem under a wide range of demand levels and CAV penetration rates, while keeping driver behavior anchored in real freeway observations and defining conflicts via physically interpretable criteria.

We fully acknowledge that this hybrid validation cannot replace large-scale field testing on real mixed CAV–HDV networks. In the revised Conclusion, we now explicitly state that 1) the current evaluation is conducted in a calibrated freeway simulation, 2) the direct transferability to urban or suburban networks is therefore limited, and 3) the lack of suitable real-world mixed-traffic datasets is a key limitation of the present work. We also clarify that, once continuous macroscopic observations of mixed CAV–HDV traffic over long freeway segments become available, we plan to retrain and evaluate MS-STGNet on those data and to systematically compare its performance with other state-of-the-art safety models in real-world environments. Below is the revised version:

Revised (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalized intersections, pedestrians, and non-motorized vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring

scalable pretraining and training strategies on larger and more diverse networks, including freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

Comment 3:

Additionally, the performance metrics used for evaluation are standard, but the manuscript does not provide a thorough analysis of the model's performance under different traffic conditions and scenarios. A more detailed sensitivity analysis and comparison with state-of-the-art models in various settings would strengthen the manuscript.

Response to Comment 3:

We appreciate this constructive suggestion. In the revised manuscript, we have clarified and strengthened the analysis of MS-STGNet under different traffic conditions and scenarios, and its comparison with state-of-the-art baselines. First, Section 6.1 (Data preparation) clearly explains that the simulation covers 500 hours of mixed CAV–HDV traffic with three representative demand levels (low, medium, high traffic volume) and five CAV penetration rates (10%, 30%, 50%, 70%, 90%), which define the set of operating scenarios used throughout the experiments (See Page 14 Line 10—16). Section 6.5 has been updated to report all metrics in Table 4 as mean ± standard deviation over five independent runs with different random seeds, and to note that the improvements of MS-STGNet over the baseline models are statistically significant at the 5% level ($p < 0.05$), thereby quantifying cross-run stability in a scenario-wise comparison.

Second, to provide a more detailed sensitivity and scenario analysis, we have explicitly highlighted several complementary results: 1) Section 6.8 (Posterior probability analyses) examines how the distributions of predicted conflict probabilities evolve with penetration rate for MS-STGNet versus STGCN and STGAT, showing that MS-STGNet maintains better separation between conflict and non-conflict classes as class imbalance increases. 2) Section 6.10, together with the new Appendix C, investigates the impact of traffic volume and speed dispersion on conflict risk and model behavior, using trajectory plots at low/medium/high volumes to illustrate how higher demand intensifies speed oscillations and conflicts, and how MS-STGNet aligns its risk predictions with these patterns more effectively than the baselines. 3) Section 6.6 now includes a computational cost analysis (GPU memory usage and parameter counts) for all deep learning baselines and MS-STGNet, complementing the accuracy-based comparison with an efficiency perspective. Taken together, these additions provide a more thorough sensitivity analysis across penetration rates, demand levels, disturbance patterns, and computational cost in line with the reviewer's recommendation. Below is the revised version:

Revised:

In Section 6.2 (Page 14 Line 34—35):

To reduce the impact of randomness and evaluate the stability of each method, all models

are trained and evaluated five times with different random seeds orders.

In Section 6.5 (Page 16 Line 38—43, Page 17, Page 18 Line 1—43):

To further assess cross-run stability, each entry in Table 4 is reported as the mean ± standard deviation over five independent runs with different random seeds. Statistical tests across the five independent runs show that the improvements of all reported metrics and penetration-rate scenarios are statistically significant at the 5% level ($p < 0.05$).

Traffic conflict prediction remains a significant challenge, particularly in distinguishing between non-conflict and conflict states. Traditional machine learning algorithms, such as SVM and XGBoost, struggle with this task compared to deep learning approaches. For example, under a 30% penetration rate, the recall rates of SVM and XGBoost were 23% and 17.8% lower, respectively, than those of the proposed MS-STGNet. Additionally, their false alarm rates increased by 27.9% and 20.0%, AUC values decreased by 24.3% and 18.9%, and accuracy was reduced by 26.4% and 20.5%. These results emphasize the importance of extracting nonlinear correlations for traffic conflict prediction.

The introduction of deep learning methods significantly improved model performance. CNN and LSTM-CNN outperformed SVM and XGBoost across all metrics, demonstrating the importance of capturing spatial dependencies and temporal correlations in conflict prediction. However, deep learning methods relying on CNNs to capture spatial dependencies face a notable limitation: they cannot model spatial similarities in unconnected grid fields. This highlights the advantage of leveraging graph neural networks (GNNs), such as STGCN and STGAT, to model semantic spatial dependencies, further enhancing performance. For instance, under a 30% penetration rate, STGAT and STGCN improved recall rates by 4.0% and 3.9%, reduced false alarm rates by 2.4% and 3.0%, increased AUC values by 2.9% and 1.5%, and improved accuracy by 4.4% and 3.2%, respectively, compared to LSTM-CNN. These results underscore the advanced capability of utilizing the inherent graph structure of road networks to extract spatial dependencies related to conflict risks. GNNs are particularly well-suited for capturing complex relationships between road segments, integrating heterogeneous road features, and learning network-wide patterns while retaining local details. Comparatively, GAT-based models often outperform GCN models by incorporating predefined adjacency matrices embedded with spatial proximity and contextual similarity, better representing spatial dependencies.

Building on prior advancements in graph-based models, the proposed MS-STGNet model demonstrated robust performance across all penetration rate scenarios. For instance, under a 50% penetration rate, MS-STGNet outperformed the next-best models by 4.9% in recall, reduced false alarm rates by 3.3%, improved AUC by 5.3%, and increased accuracy by 3.3%. Notably, as shown in Table 4, MS-STGNet achieved a significant reduction in false alarm rates, with improvements of 23.9%, 24.0%, and 23.8% under 50%, 70%, and 90% penetration rates, respectively. This improvement can be attributed to the manifold similarity module, which reduces misjudgments in conflict-prone areas of traffic flow—a point further analyzed in subsequent sections.

Because the task is a binary conflict/non-conflict prediction problem on a large-scale dataset, the standard deviations across runs are generally small for all models. Nevertheless, the reported mean ± standard deviation helps to reveal relative robustness: MS-STGNet

maintains consistent advantages over STGCN and STGAT across different penetration rates, and in most cases exhibits comparable or slightly lower variation in key metrics. This indicates that the improvements of MS-STGNet are not due to a single favourable initialization but are reproducible under different random seeds.

Table 4
Performance of Different Models on Datasets.

Penetration rates	Metric	SVM	XGBoost	CNN	LSTM-CNN	STGCN	STGAT	MS-STGNet
10%	Recall	0.531 ± 0.049	0.577 ± 0.037	0.713 ± 0.031	0.726 ± 0.024	0.766 ± 0.011	0.782 ± 0.019	**0.797**↑1.92% ± 0.014
	False alarm rate	0.440 ± 0.047	0.413 ± 0.038	0.206 ± 0.029	0.201 ± 0.017	0.175 ± 0.012	0.165 ± 0.009	**0.150**↑9.09% ± 0.009
	AUC	0.588 ± 0.049	0.632 ± 0.039	0.758 ± 0.034	0.769 ± 0.021	0.790 ± 0.020	0.807 ± 0.024	**0.824**↑2.11% ± 0.016
	Accuracy	0.581 ± 0.039	0.652 ± 0.055	0.788 ± 0.029	0.803 ± 0.030	0.830 ± 0.014	0.829 ± 0.017	**0.855**↑3.01% ± 0.011
	G-mean	0.543 ± 0.067	0.581 ± 0.053	0.745 ± 0.037	0.769 ± 0.026	0.793 ± 0.021	0.803 ± 0.016	**0.820**↑2.12% ± 0.017
30%	Recall	0.578 ± 0.040	0.630 ± 0.036	0.742 ± 0.022	0.738 ± 0.028	0.777 ± 0.016	0.778 ± 0.013	**0.808**↑3.86% ± 0.010
	False alarm rate	0.417 ± 0.049	0.338 ± 0.054	0.194 ± 0.024	0.173 ± 0.013	0.143 ± 0.013	0.149 ± 0.015	**0.138**↑3.50% ± 0.013
	AUC	0.596 ± 0.047	0.650 ± 0.034	0.757 ± 0.027	0.781 ± 0.025	0.796 ± 0.020	0.810 ± 0.017	**0.839**↑3.58% ± 0.007
	Accuracy	0.592 ± 0.065	0.651 ± 0.048	0.781 ± 0.033	0.789 ± 0.018	0.821 ± 0.023	0.833 ± 0.021	**0.856**↑2.76% ± 0.015
	G-mean	0.592 ± 0.041	0.644 ± 0.046	0.773 ± 0.039	0.775 ± 0.020	0.815 ± 0.014	0.816 ± 0.015	**0.831**↑1.84% ± 0.012
50%	Recall	0.564 ± 0.047	0.593 ± 0.054	0.767 ± 0.028	0.788 ± 0.031	0.803 ± 0.023	0.828 ± 0.019	**0.877**↑5.92% ± 0.009
	False alarm rate	0.417 ± 0.047	0.332 ± 0.039	0.181 ± 0.021	0.165 ± 0.019	0.139 ± 0.018	0.138 ± 0.015	**0.105**↑23.91% ± 0.017
	AUC	0.580 ± 0.044	0.672 ± 0.042	0.790 ± 0.036	0.802 ± 0.026	0.823 ± 0.022	0.833 ± 0.024	**0.886**↑6.36% ± 0.013
	Accuracy	0.590 ± 0.035	0.650 ± 0.053	0.794 ± 0.032	0.830 ± 0.015	0.852 ± 0.019	0.857 ± 0.016	**0.890**↑3.85% ± 0.008
	G-mean	0.563 ± 0.050	0.642 ± 0.045	0.789 ± 0.027	0.813 ± 0.030	0.826 ± 0.014	0.843 ± 0.011	**0.887**↑5.22% ± 0.011
70%	Recall	0.571 ± 0.038	0.617 ± 0.051	0.759 ± 0.025	0.749 ± 0.028	0.770 ± 0.014	0.789 ± 0.018	**0.816**↑3.42% ± 0.012
	False alarm rate	0.427 ± 0.050	0.329 ± 0.047	0.168 ± 0.030	0.170 ± 0.022	0.141 ± 0.017	0.125 ± 0.012	**0.095**↑24.00% ± 0.010
	AUC	0.576 ± 0.037	0.673 ± 0.050	0.782 ± 0.023	0.781 ± 0.029	0.816 ± 0.020	0.822 ± 0.015	**0.860**↑4.62% ± 0.015
	Accuracy	0.589 ± 0.046	0.668 ± 0.057	0.809 ± 0.035	0.801 ± 0.011	0.830 ± 0.013	0.836 ± 0.020	**0.898**↑7.42% ± 0.007
	G-mean	0.588 ± 0.051	0.639 ± 0.045	0.802 ± 0.028	0.795 ± 0.027	0.811 ± 0.022	0.828 ± 0.009	**0.860**↑3.86% ± 0.014
90%	Recall	0.597 ± 0.053	0.622 ± 0.050	0.782 ± 0.034	0.770 ± 0.021	0.783 ± 0.023	0.809 ± 0.010	**0.819**↑1.24% ± 0.008
	False alarm rate	0.388 ± 0.038	0.352 ± 0.041	0.170 ± 0.024	0.147 ± 0.027	0.130 ± 0.016	0.122 ± 0.022	**0.093**↑23.77% ± 0.011
	AUC	0.595 ± 0.040	0.658 ± 0.031	0.793 ± 0.038	0.786 ± 0.012	0.821 ± 0.011	0.832 ± 0.013	**0.860**↑3.37% ± 0.009
	Accuracy	0.591 ± 0.063	0.682 ± 0.053	0.812 ± 0.020	0.835 ± 0.024	0.857 ± 0.028	0.873 ± 0.018	**0.896**↑2.63% ± 0.016
	G-mean	0.600 ± 0.048	0.635 ± 0.035	0.798 ± 0.031	0.802 ± 0.018	0.822 ± 0.019	0.839 ± 0.025	**0.863**↑2.86% ± 0.013

In Section 6.10 (Page 23 Line 28—29, Page 24 Line 1—17):

Beyond penetration rates, we also examined how traffic volume and the resulting speed dispersion patterns affect conflict risk and model behavior. In the simulation, different representative demand levels were considered over a total of 500 hours, covering low-,

medium-, and high-volume conditions. The supplementary trajectory plots in Appendix C (Figures C.1–C.3) show that as traffic volume increases, pronounced speed oscillations emerge along the segment and become more frequent and severe. This indicates that, even under mixed CAV–HDV conditions, higher demand intensifies vehicle interactions and amplifies the likelihood of conflicts, which supports our use of traffic state variations as predictors of conflict occurrence. A closer inspection of these trajectories further highlights the role of different vehicle classes and CAV penetration as key traffic features. The green and blue trajectories representing HDVs exhibit larger amplitude and higher-frequency speed fluctuations than the red trajectories representing CAVs, reflecting more aggressive driving behavior and delayed responses in the human-driven fleet. Heavy vehicles (trucks) introduce additional instability due to their limited acceleration and deceleration capabilities and larger size, which force surrounding vehicles to adjust their speeds more frequently and create pronounced perturbation zones. As CAV penetration increases, these unstable zones shrink and the gaps between high-speed and low-speed vehicle clusters are gradually bridged by heterogeneous CACC queues, leading to smoother trajectories and reduced speed dispersion. Combined with the segment-level risk profiles in Fig.9, these observations indicate that CAV penetration rate, traffic volume, and the resulting speed separation patterns are among the most influential traffic features for conflict prediction in the proposed framework: MS-STGNet is particularly effective at aligning its predicted risk with these underlying speed dispersion structures, while STGCN and STGAT tend to generate spurious conflict probabilities in disturbance zones.

In Appendix C (Page 28 Line 7—13, Page 29):

Appendix C. Supplementary vehicle position–speed trajectories

In this appendix, we provide additional vehicle position–speed trajectory plots for three representative demand levels, corresponding to low-, medium-, and high-volume conditions. For each traffic volume, the trajectories are shown separately for the pre-merging, merging, and post-merging segments, with different colors indicating HDVs, CAVs, and heavy vehicles (trucks). These plots illustrate how increasing traffic volume and changes in vehicle composition lead to more pronounced speed oscillations and perturbation zones, complementing the case study around Fig. 9 in the main text and supporting the discussion in Section 6.10 on the impact of traffic volume, CAV penetration, and speed separation on conflict risk.

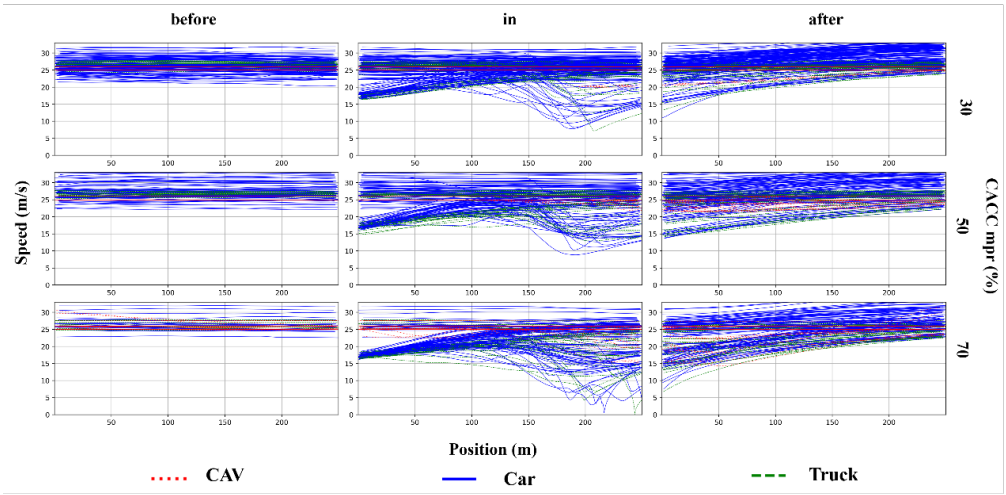

Fig. C1. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 3000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

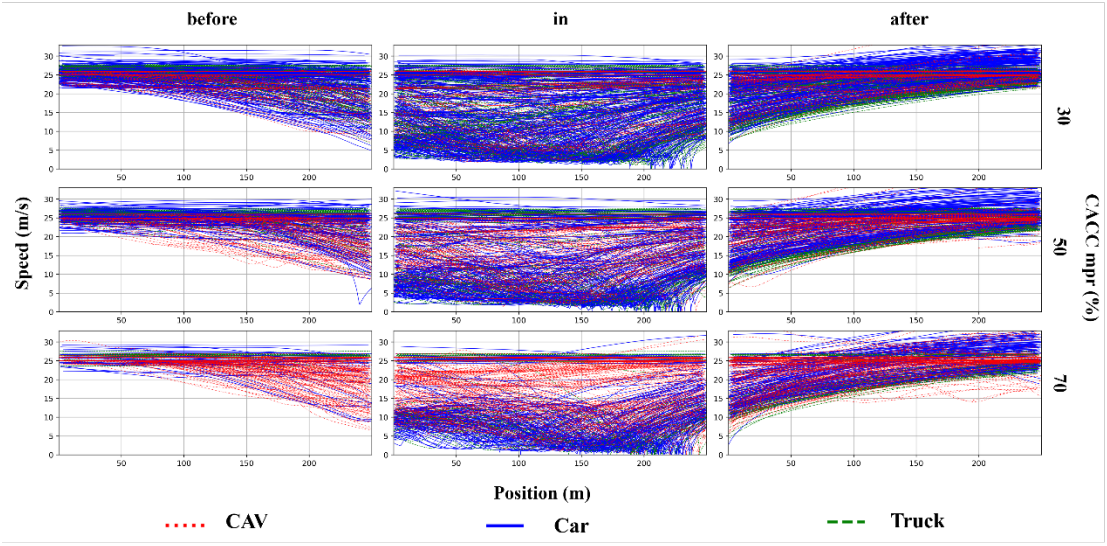

Fig. C2. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 6000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

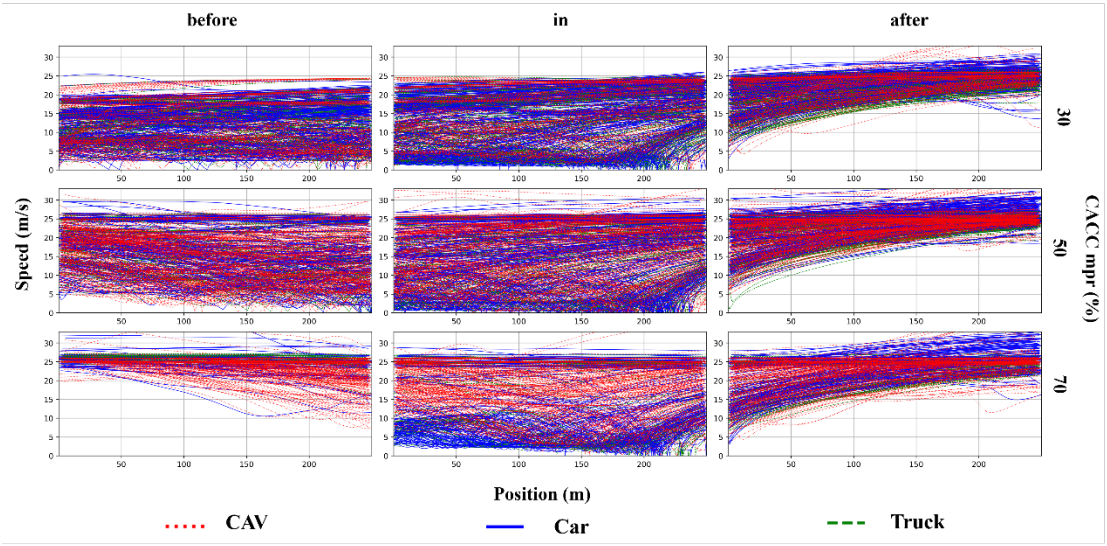

Fig. C3. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 9000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

In Section 7 (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway

segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalised intersections, pedestrians, and non-motorised vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring scalable pretraining and training strategies on larger and more diverse networks, including freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

In Section 6.6 (Page 18 Line 44—52, Page 19 Line 1—16):

In real-world deployment, predictive accuracy is the primary requirement for traffic safety applications, while the hardware cost of the deployed model constitutes a secondary but still crucial consideration for practical implementation. To highlight the computational overhead of different approaches, Table 5 reports three indicators under five CAV penetration-rate scenarios: GPU-MUT (peak GPU memory usage during training), GPU-MUI (peak GPU memory usage during inference), and the number of trainable parameters. For the classical machine-learning baselines (SVM and XGBoost), GPU-based indicators are omitted ("–") because they are trained and executed on CPU and their memory footprint is negligible compared with deep models in our setting.

Several observations can be made from Table 5. First, among the deep learning baselines, STGCN consistently has the largest parameter count and highest GPU memory usage, with STGAT slightly smaller but still noticeably heavier than CNN and LSTM-CNN. For example, at a 50% penetration rate, STGCN and STGAT require 479,816 and 426,572 parameters, respectively, and their GPU-MUT values reach 4,497 MiB and 4,681 MiB. By contrast, the proposed MS-STGNet uses fewer parameters than both graph-based baselines (395,428 at 50% penetration) and reduces peak GPU memory by roughly 10–15% in training (e.g., 4,059 MiB versus 4,497–4,681 MiB) and 15–25% in inference (e.g., 2,710 MiB versus 3,216–3,587 MiB), while still incorporating a manifold-similarity module and adaptive fusion. Compared with CNN and LSTM-CNN, MS-STGNet understandably incurs moderately higher GPU memory usage due to the additional graph operations, but remains in the same order of magnitude and does not introduce prohibitive overhead.

Table 5
The computational performance of different models on dataset.

Penetration rates	Metric	SVM	XGBoost	CNN	LSTM-CNN	STGCN	STGAT	MS-STGNet
10%	GPU-MUT	—	—	4,333MiB	4,443MiB	5,574MiB	5,802MiB	**5,031MiB**
	GPU-MUI	—	—	2,283MiB	2,799MiB	4,446MiB	3,986MiB	**3,359MiB**
	Parameters	—	—	298,742	346,251	594,758	528,759	**490,154**
30%	GPU-MUT	—	—	4,419MiB	4,530MiB	5,684MiB	5,917MiB	**5,130MiB**
	GPU-MUI	—	—	2,328MiB	2,854MiB	4,534MiB	4,065MiB	**3,425MiB**
	Parameters	—	—	304,621	353,064	606,462	539,164	**499,800**
50%	GPU-MUT	—	—	3,496MiB	3,584MiB	4,497MiB	4,681MiB	**4,059MiB**
	GPU-MUI	—	—	1,842MiB	2,258MiB	3,587MiB	3,216MiB	**2,710MiB**
	Parameters	—	—	241,008	279,335	479,816	426,572	**395,428**
70%	GPU-MUT	—	—	3,085MiB	3,162MiB	3,968MiB	4,130MiB	**3,581MiB**
	GPU-MUI	—	—	1,625MiB	1,992MiB	3,165MiB	2,838MiB	**2,391MiB**
	Parameters	—	—	212,656	246,474	423,370	376,390	**348,909**
90%	GPU-MUT	—	—	2,983MiB	3,058MiB	3,837MiB	3,994MiB	**3,463MiB**
	GPU-MUI	—	—	1,572MiB	1,927MiB	3,061MiB	2,744MiB	**2,312MiB**
	Parameters	—	—	205,636	238,338	409,394	363,965	**337,392**

Overall, these results indicate that MS-STGNet achieves superior predictive performance (as shown in Table 4) with a computational cost that is only modestly higher than conventional CNN-based models and clearly lower than that of STGCN and STGAT. This suggests that the proposed architecture strikes a reasonable balance between accuracy and efficiency, making it suitable for deployment in practical mixed CAV–HDV conflict prediction systems. We do not report wall-clock training or inference time, as such measurements are highly dependent on specific hardware, software environments, and background system load; instead, we focus on parameter counts and GPU memory usage, which provide hardware-agnostic indicators of computational complexity.

Comment 4:

Lastly, the manuscript could benefit from a more in-depth discussion of the potential limitations and future work. The proposed framework, while showing promising results, may face challenges in real-time implementation and scalability, which should be addressed in the manuscript.

Response to Comment 4:

We thank the reviewer for this helpful suggestion. In the revised manuscript, we have expanded the Conclusion section (Section 7) to more explicitly discuss the limitations and future work, including issues related to real-time implementation and scalability. Specifically, we now 1) clarify the scope of our current evaluation (a simulated 14 km multilane freeway calibrated on highD data, without urban/suburban networks or multimodal road users), 2) acknowledge methodological constraints such as the use of a static manifold-similarity matrix and a binary conflict/non-conflict output, and 3) outline concrete future directions, including validation on emerging mixed CAV–HDV field datasets, extension to more diverse networks

and contextual variables, development of online/adaptive manifold learning, and graded conflict-severity modelling. These additions better delimit the applicability of the present framework and directly address the reviewer's concerns about its potential real-time deployment and scalability. Below is the revised version:

Revised (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalised intersections, pedestrians, and non-motorized vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring scalable pretraining and training strategies on larger and more diverse networks, including freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

Comments from Reviewer #2:

Comment 1:

Please provide specific recent (<5-year) prior works supporting Section 2.3.

Response to Comment 1:

Thank you for this helpful suggestion.In the revised manuscript, we have added several recent references that explicitly exploit manifold-based representations for traffic state modelling and traffic safety analysis. Specifically, in Section 2.3, we now discuss:

- Su et al. (2020), who use a convolutional variational auto-encoder to extract low-dimensional manifold representations of daily urban traffic flow for clustering;
- Seo (2023), who applies Uniform Manifold Approximation and Projection (UMAP), a manifold-learning-based dimension reduction method, to visualize large-scale network traffic states and identify distinct congestion regimes;
- Liu et al. (2022), who incorporate manifold characteristics of traffic flow into a transfer-learning-based highway crash risk evaluation model and show that manifold features improve the discrimination between high- and low-risk traffic states.

[Liu, Q., Li, C., Jiang, H., Nie, S., & Chen, L. (2022). Transfer learning-based highway crash risk evaluation considering manifold characteristics of traffic flow. *Accident Analysis & Prevention*, *168*, 106598.

Seoa, T. (2023). Understanding large-scale traffic flow using model-based and data-driven dimension reduction: with COVID-19 and Olympic-Paralympic case study. *EU Science Hub*, 124.

Su, M. T., Zheng, J., & Zhang, Z. P. (2020). Clustering Mining of Urban Traffic Flow Based on CVAE. *Journal of Traffic and Logistics Engineering Vol*, *8*(2).]

These additions strengthen the motivation of Section 2.3 and provide up-to-date support for the use of manifold similarity in our proposed MS-STGNet framework. Below is the revised version:

Revised (Page 5 Line 13—23):

Recent studies have begun to explicitly model traffic flow on low-dimensional manifolds. For example, Su et al. (2020) used a convolutional variational auto-encoder to extract low-dimensional manifold representations of daily urban traffic flow and showed that clustering in this latent space reveals meaningful traffic patterns. Seoa (2023) applied Uniform Manifold Approximation and Projection (UMAP), a non-linear dimension-reduction method based on manifold learning, to obtain two-dimensional embeddings of large-scale network traffic states, demonstrating that the learned manifold coordinates intuitively capture different congestion regimes. In the field of traffic safety, Liu et al. (2022) incorporated manifold characteristics of traffic flow into a transfer-learning-based highway crash risk evaluation model and reported improved discrimination between high- and low-risk traffic states compared with models that rely solely on Euclidean features. These studies indicate that manifold-based representations can provide a more faithful description of the dynamic evolution and similarity of traffic systems than conventional distance measures in the original feature space.

Comment 2:

Does Fig. 8 present only a subset of the complete vehicle position-speed trajectories corresponding to the silhouette in Fig. 1? Please specify how lane position is defined and measured (e.g., lane index vs. lateral offset in meters)?

Response to Comment 2:

Thank you very much for this helpful comment. In the revised version, we have clarified in Section 6.9 that Fig. 8 shows the vehicle position–speed trajectories for a 250 m subsegment of the on-ramp merging area depicted in Fig. 1, selected because this location exhibits the most pronounced speed oscillations under high CAV penetration and thus better illustrates the speed separation between CAVs and HDVs. We also now explicitly state that the lateral axis in Fig. 8 represents the lateral offset in meters (rather than a discrete lane index), measured across the cross-section of the entire roadway. Since our analysis focuses on speed disturbances within the CAV and HDV systems rather than lane-by-lane differences, we deliberately avoid lane coloring and instead emphasize the contrast in speed fluctuation patterns between the two vehicle groups. Below is the revised version:

Revised (Page 23 Line 5—8):

We selected a segment of approximately 250 meters of an on-ramp merging scenario to illustrate the position-velocity trajectories of vehicles from both the HDV and CAV groups (as shown in Fig.8). Compared to the main highway, the merging scenario on the ramp exhibits more pronounced fluctuations and oscillations in vehicle speed, which facilitates a clearer observation of the differences between the two groups.

Comment 3:

The formula (15) seems wrong. The ReLU takes a single tensor as input, the comma notation is non-standard and ambiguous.

Response to Comment 3:

We thank the reviewer for pointing out the issue in Eq. (15). In the original manuscript, ReLU was mistakenly written as $\mathrm{ReLU}(\mathbf{M}_{\mathrm{lt}}, \mathbf{M}_{\mathrm{rt}})$ which is non-standard and indeed ambiguous, since ReLU should take a single tensor as input. We have revised the notation in the manuscript accordingly to avoid ambiguity, and in response to comments from other reviewers, this formula has been moved to Appendix B. Below is the revised version:

Revised (Page 28 Appendix B, B.5):

$$\widetilde{\mathbf{A}^{\star}} = \mathbf{I_N} + \mathrm{softmax}\big(\mathrm{ReLU}(\mathbf{M}_{\mathrm{lt}}\mathbf{M}_{\mathrm{rt}})\big)$$

Comment 4:

In formula (16), the placement of b_k outside the summation is confusing.

Response to Comment 4:

We appreciate the reviewer's insightful comment regarding the placement of b_k in Eq. (16). In the original manuscript, the bias term b_k was written outside the summation while still indexed by k, which is indeed confusing and mathematically ambiguous. As correctly pointed out by the reviewer, if the bias depends on k, it should appear inside the summation. In our implementation, each order has its own learnable bias associated with the corresponding weights. We have revised the notation in the manuscript accordingly to avoid ambiguity. After the article structure was readjusted, it is now Eq. (11). Below is the revised version:

Revised (Page 12 Eq.11):

$$\mathbf{Z}_{\mathbf{t}}^{*} = \sum_{k=1}^{K} \left((\mathbf{P}_{\mathbf{f}}^{*})^{k} \mathbf{X}^{\mathbf{t}} \mathbf{W}_{\mathbf{k},1} + (\mathbf{P}_{\mathbf{b}}^{*})^{k} \mathbf{X}^{\mathbf{t}} \mathbf{W}_{\mathbf{k},2} + \widetilde{\mathbf{A}_{\star}} \mathbf{X}^{\mathbf{t}} \mathbf{W}_{\mathbf{k},3} + \mathbf{b}_{\mathbf{k}} \right)$$

Comment 5:

There are several types of f in Eq. (18) and Eq. (19). What's the difference?

Response to Comment 5:

We thank the reviewer for pointing out the ambiguity regarding the different notations of f in Eqs. (18) and (19). In our model, all these symbols refer to learnable 1D convolution kernels, but they play slightly different roles:

- In Eq. (18), $\mathbf{f}^{l,k} \in \mathbb{R}^{C}$ denotes the generic 1D convolution kernel of the l-th TCN layer and the k-th output channel, and $\mathbf{f}^{l,k}(m)$ is its m-th coefficient. This equation defines the general form of a dilated causal convolution.
- In Eq. (19), $\mathbf{f}_k^{(0)}$ and $\mathbf{f}_k^{(1)}$ denote the convolution kernels used in the first and second dilated convolution within the k-th residual TCN block, respectively. They are specific instances of the generic kernel used in Eq. (18), and we use the superscripts (0) and (1) to distinguish the two convolutional layers inside a block.

To avoid confusion, we have revised the manuscript to explicitly clarify these roles. In particular, we now 1) add an explicit description of $\mathbf{f}^{l,k}$ below Eq. (18), and 2) clarify below Eq. (19) that $\mathbf{f}_k^{(0)}$ and $\mathbf{f}_k^{(1)}$ are the kernels of the two dilated convolutions in the k-th residual TCN block. We believe this resolves the ambiguity about the different types of f in these equations. Below is the revised version:

Revised (Page 12 Line 38—39, Page 13 Line 2—4):

where C is the number of channels; d is the dilation factor; m indexes the dilation intervals; and $\mathbf{f}^{l,k} \in \mathbb{R}^{C}$ denotes the 1D convolution kernel of the l-th TCN layer and the k-th output channel.

where $\mathbf{f}_k^{(0)}$ and $\mathbf{f}_k^{(1)}$ are also 1D convolution kernels, corresponding to the first and second dilated convolutions in the k-th residual TCN block, respectively. They are specific instances of the generic kernel $\mathbf{f}^{l,k}$ defined in Eq. (18), but we use superscripts (0) and (1) to

distinguish the two convolutional layers within each block;

Comment 6:

In formula (24), the definitions of L_{LDAM} *and* Δ_y *are set consecutively without a separator, which can be misread as a single expression.*

Response to Comment 6:

We appreciate the reviewer's comment regarding the readability of Eq. (24). In the original manuscript, the definitions of $\mathcal{L}_{\mathrm{LDAM}}$ and Δ_y were typeset consecutively in the same display without any separator, which could indeed be misread as a single expression. To avoid this ambiguity, we have revised Eq. (24) to clearly separate the definitions. After the article structure was readjusted, it is now Eq. (19). Below is the revised version:

Revised (Page 13 Eq.19):

$$\begin{cases} \mathcal{L}_{Focal} = -\alpha_t (1-p_t)^{\gamma} \log(p_t) \\ \mathcal{L}_{LDAM} = -\log \dfrac{\exp(z_y - \Delta_y)}{\exp(z_y - \Delta_y) + \sum_{j \neq y} \exp(z_j)}, \quad \Delta_y = \dfrac{S}{n_y^{\sigma}} \\ \mathrm{Loss}(\mathbf{Y}, \widehat{\mathbf{Y}}) = \alpha \cdot \mathcal{L}_{LDAM} + \beta \cdot \mathcal{L}_{Focal} \end{cases}$$

Comments from Reviewer #3:

Comment 1:

The idea of using a manifold-based similarity to construct adaptive graphs is well motivated and implemented cleanly. However, the formulation is conceptually close to existing adaptive adjacency or attention mechanisms in STGAT-type models. A clearer discussion of how the manifold metric differs in principle or in computational benefit would strengthen the methodological message. I think, this is the weakest point, also being positioned as a selling point.

Response to Comment 1:

We sincerely thank the reviewer for these thoughtful comments on the novelty of MS-STGNet and its relation to existing manifold-based traffic models and STGAT-type adaptive graph networks.

In the revised manuscript, we have clarified our contributions at both the problem and method levels. From the problem perspective, we emphasize that our primary goal is real-time conflict prediction in mixed CAV–HDV freeway traffic, a setting where existing work is still limited. To ensure robustness in this new safety-critical application, we deliberately build on mature components (residual CNN, TCN, spatiotemporal GNNs) while introducing a manifold-similarity graph as a physically meaningful prior rather than proposing an entirely new architecture for its own sake.

From the methodological perspective, we now explicitly distinguish our approach from prior manifold-learning studies and STGAT-type adaptive adjacency mechanisms. Section 2.3 has been expanded to include recent manifold-based traffic-flow and safety studies and to clarify that these works mainly use manifold embeddings for clustering, visualization, or as features in conventional models, without embedding manifold-based traffic-state similarity into a spatiotemporal GNN for online conflict prediction. In contrast, MS-STGNet integrates a pre-computed manifold similarity matrix, derived from historical traffic states, as an interpretable prior that constrains adaptive adjacency learning for mixed CAV–HDV conflicts.

We also revise Section 2.2 and Section 5.3.1 to clearly contrast our manifold-similarity graph with standard STGAT mechanisms: instead of learning adjacency solely from instantaneous node features at each time step, MS-STGNet initializes the graph from manifold distances computed offline and then performs lightweight adaptive refinement. This design ties the learned graph to physically meaningful traffic-state geometry while keeping the per-iteration computational cost comparable to standard STGNNs. Finally, in Section 6.5 we explicitly highlight that the manifold-similarity prior contributes to the reduction of false alarm rates and improved robustness compared with STGCN and STGAT, especially at medium-to-high CAV penetration rates.

We hope these revisions and clarifications make the unique contributions and methodological positioning of MS-STGNet more evident. Below is the revised version:

Revised:

In Section 1 (Page 2 Line 45—49, Page 3 Line 10—14):

Second, we propose MS-STGNet, a spatiotemporal graph neural network that fuses

physical adjacency and semantic features for traffic conflict prediction in mixed CAV–HDV traffic. The framework intentionally builds on mature components (e.g., residual CNN and TCN) to ensure robustness in this new application setting, while introducing a manifold-similarity graph as a physically meaningful prior for adaptive adjacency, which has not been explored in existing mixed-traffic conflict prediction models.

In MS-STGNet, a manifold similarity graph module has been developed and implemented. By leveraging a similarity matrix derived from traffic state data within the manifold space, we provide prior knowledge regarding the evolution of traffic states. The manifold-similarity module incorporates a broader array of traffic-flow attributes during neighbor selection and uses a pre-computed manifold similarity matrix as an interpretable structural prior, thereby reducing the propensity for false-positive conflict-event predictions.

In Section 2.2 (Page 4 Line 41—44):

In addition, existing spatiotemporal graph-based safety models typically define spatial dependencies through fixed adjacency matrices or adaptive attention mechanisms in the original feature space, and rarely exploit manifold-based traffic-state similarity as an explicit prior, particularly in mixed CAV–HDV traffic environments.

In Section 2.3 (Page 5 Line 26—29):

Additionally, few studies have attempted to integrate the concept of state transitions in manifold learning into deep learning frameworks, and, to the best of our knowledge, none has embedded manifold-based traffic-state similarity into a spatiotemporal graph neural network for real-time conflict prediction in mixed CAV–HDV traffic.

In Section 5.3.1 (Page 12 Line 1—10):

Conceptually, the proposed manifold-similarity graph plays a role that is related to, but distinct from, the adaptive adjacency mechanisms used in STGAT-type models. In conventional STGAT, edge weights are learned solely from instantaneous node features via attention, and the adjacency matrix is dynamically reconstructed at each time step. In MS-STGNet, the adjacency structure is instead initialized from manifold distances computed over historical traffic states, which encode long-term traffic-flow evolution and physically meaningful similarity between spatiotemporal patterns. The subsequent adaptive update in MSGNet refines this manifold-based prior rather than discarding it. This separation between a manifold-informed prior graph that reflects the geometric structure of traffic dynamics and a lightweight adaptive refinement brings two benefits: it constrains the learned graph to remain consistent with empirical traffic-state geometry, and it limits the additional per-iteration cost compared with fully attention-based dynamic graphs, keeping the overall complexity comparable to that of standard STGNN models.

In Section 6.5 (Page 18 Line 31—43):

These empirical results also clarify how MS-STGNet differs in practice from STGAT-type adaptive graph models. Although both approaches employ graph-based representations, STGAT relies on feature-driven attention to construct adjacency at each time step, which can be sensitive to local fluctuations in highly imbalanced conflict datasets. By contrast, MS-

STGNet constrains the adaptive graph updates within a manifold-similarity prior derived from historical traffic states. As the market penetration of CAVs increases and pronounced speed separation emerges, this manifold-informed prior helps the model avoid spuriously high conflict probabilities in non-conflict regions, leading to consistently lower false alarm rates and more stable performance across all penetration scenarios. In this sense, our findings are consistent with previous studies showing that graph-based spatiotemporal models such as STGCN and STGAT outperform traditional machine-learning and sequence models in traffic prediction tasks, while further extending them by explicitly incorporating a manifold-based state similarity prior into the adaptive graph learning process. At the same time, our results complement recent manifold-learning approaches for traffic state analysis by demonstrating that manifold-informed similarity can be embedded into deep spatiotemporal graph networks to improve conflict prediction in mixed CAV–HDV freeway traffic.

Comment 2:

The simulation framework and evaluation are comprehensive, very solid! The ablation study (Section 6.6) is exemplary and demonstrates the incremental gain from each module. Nevertheless, the validation remains limited to simulated data; any test on partially real or hybrid datasets (like MITRA, HDSim, etc.) would raise the practical impact substantially.

Response to Comment 2:

We sincerely appreciate this important comment. At the early stage of this study, our initial intention was indeed to develop and validate MS-STGNet directly on real-world mixed CAV–HDV data. However, after a thorough review of existing datasets, we found that currently available data sources cannot simultaneously meet the two core requirements of our problem: 1) truly mixed CAV–HDV traffic, and 2) long, spatially continuous freeway segments with macroscopic measurements (flow, speed, occupancy) suitable for segment-level conflict prediction over continuous time series.

On the one hand, classical trajectory datasets such as NGSIM and highD contain only human-driven vehicles and therefore do not match our target mixed-traffic scenario. Nevertheless, to ensure that our simulation is not based on purely theoretical assumptions, we calibrated the human-driven car-following model directly on highD freeway trajectories, so that HDV behaviour in the simulation reflects empirically observed acceleration, deceleration, and headway patterns rather than arbitrary parameter choices.

With respect to the specific datasets mentioned by the reviewer, we have carefully examined their suitability for our study:

MITRA is a high-resolution drone-based trajectory dataset on a 900m urban freeway segment in Milan, covering all traffic states with ramps and lane changes. It is highly valuable for microscopic analysis of human-driven behavior. However, the current release contains only HDVs (no explicit CAV logic). Thus, MiTra is well suited for refining microscopic models but does not directly provide the mixed CAV–HDV, segment-level macroscopic data needed for our conflict prediction framework.

[Chaudhari, A. A., Treiber, M., & Okhrin, O. (2025). Mitra: A drone-based trajectory data for an all-traffic-state inclusive freeway with ramps. Scientific Data, 12(1), 1174.]

HDSim is a cognitively inspired human-like traffic simulation framework that generates realistic microscopic scenario for testing autonomous driving systems. Conceptually, it is closer to our own SUMO/Plexe-based simulator than to a ready-made macroscopic dataset: it focuses on scene-level microscopic interaction rather than delivering continuous freeway-level traffic-state time series with conflict labels. Using HDSim would therefore require porting our entire controller, detection, surrogate safety (TTC/DRAC/DDR), and aggregation pipeline to another engine, which we regard as valuable for future cross-simulator robustness studies but beyond the scope of the present work.

[Li, W., Wu, H., Gao, H., Mao, B., Xu, F., & Zhong, S. (2025). LLM-based Human-like Traffic Simulation for Self-driving Tests. arXiv preprint arXiv:2508.16962.]

On the other hand, recent autonomous-vehicle datasets such as the Lyft Level 5 AV Dataset (Houston et al., 2021), nuScenes (Caesar et al., 2020), and the Waymo Open Dataset (Sun et al., 2020) do provide mixed traffic with AVs/CAVs, but their structure is not well suited to our macroscopic conflict-prediction task. As summarized in Table 1 of Hu et al. (2022), these AV datasets are organized into short trajectory segments: Waymo comprises 1,000 segments with a temporal resolution of 0.1 s and a typical segment length of 20 s; Lyft Level 5 contains 366 segments at 0.2 s resolution and 25–45 s duration; and nuScenes includes 1,000 segments with 0.5 s resolution and 20 s duration. These segments are collected from the viewpoint of individual AVs and are neither spatially contiguous along a single freeway facility nor temporally continuous over long periods. Hu et al. (2022) explicitly note that substantial preprocessing and reconstruction are required even to obtain usable car-following trajectories from these segment-based recordings, and that the resulting data remain fragmented in space and time for macroscopic analyses.

Table 1. Overview of three AV trajectory dataset.

Dataset	Number of segments	Resolution (s)	Length of each segment (s)
Waymo	1000	0.1	20
Lyft	366	0.2	25–45
nuScenes	1000	0.5	20

[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).

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., ... & Ondruska, P. (2021, October). One thousand and one hours: Self-driving motion prediction dataset. In Conference on Robot Learning (pp. 409-418). PMLR.

Hu, X., Zheng, Z., Chen, D., Zhang, X., & Sun, J. (2022). Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research. *Transportation Research Part C: Emerging Technologies*, *134*, 103490.

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).]

Similarly, Zhang et al. (2025) show that such AV trajectory datasets are particularly suitable for microscopic behaviour analysis and data-driven stochastic fundamental-diagram modelling, but they are inherently sparse in space and time and therefore not directly aligned

with macroscopic, segment-level modelling of traffic states over extended freeway sections. In our study, however, the prediction target is whether a given freeway segment will experience a conflict within a continuous time series, based on macroscopic indicators (flow, speed, occupancy) monitored along a 14 km stretch. This requires long, contiguous observations along one facility, which current AV datasets do not provide.

[Zhang, X., Yang, K., Sun, J., & Sun, J. (2025). Stochastic fundamental diagram modeling of mixed traffic flow: A data-driven approach. *Transportation Research Part C: Emerging Technologies*, *179*, 105279.]

These limitations are consistent with the broader challenges identified in recent reviews of machine-learning-based crash prediction. For example, Ali et al. (2024) point out that empirical safety studies for mixed CAV–HDV environments are still scarce, that most real-time crash and conflict prediction models are developed for conventional freeway or urban networks without CAVs, and that many studies necessarily rely on simulation or indirectly inferred data due to the lack of suitable field datasets. Against this background, we adopted a hybrid strategy that combines empirical calibration with large-scale simulation.

[Ali, Y., Hussain, F., & Haque, M. M. (2024). Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention*, *194*, 107378.]

Given these data limitations, we adopted a hybrid "real-trajectory calibration + simulation-based testing" strategy. At the microscopic level, we construct mixed CAV–HDV traffic with multiple penetration rates (10%, 30%, 50%, 70%, and 90%) and generate trajectories at 0.2s resolution. Using these trajectories, we compute widely used surrogate safety measures TTC, DRAC, and DDR—to identify both longitudinal and lateral conflicts. At the macroscopic level, we aggregate the simulated data along a 14 km four-lane freeway with ramps and extract continuous time series of flow, speed, and occupancy for each segment, paired with the conflict/non-conflict labels derived from TTC/DRAC/DDR. This design allows us to study conflict prediction as a segment-level time-series classification problem under a wide range of demand levels and CAV penetration rates, while keeping driver behavior anchored in real freeway observations and defining conflicts via physically interpretable criteria.

We fully acknowledge that this hybrid validation cannot replace large-scale field testing on real mixed CAV–HDV networks. In the revised Conclusion, we now explicitly state that 1) the current evaluation is conducted in a calibrated freeway simulation, 2) the direct transferability to urban or suburban networks is therefore limited, and 3) the lack of suitable real-world mixed-traffic datasets is a key limitation of the present work. We also clarify that, once continuous macroscopic observations of mixed CAV–HDV traffic over long freeway segments become available, we plan to retrain and evaluate MS-STGNet on those data and to systematically compare its performance with other state-of-the-art safety models in real-world environments. Below is the revised version:

Revised (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional

expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalized intersections, pedestrians, and non-motorized vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring scalable pretraining and training strategies on larger and more diverse networks, including freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

Comment 3:

Improvements over STGCN and STGAT are numerically evident but modest. Reporting standard deviations or statistical significance across runs would help to substantiate the claimed stability.

Response to Comment 3:

We thank the reviewer for this helpful comment. In the original manuscript, our notion of "stability" mainly referred to performance consistency across different CAV penetration rates, as highlighted by the persistent reduction in false alarm rates compared with STGCN and STGAT (See Page 18 Line 21—24). We agree that cross-run variability and statistical significance should also be evaluated.

In the revised version, we retrain all models five times with different random seeds. Section 6.2 has been updated to describe this protocol, and Table 4 now reports each metric as mean ± standard deviation over these five runs. Because the task is a binary conflict/non-conflict prediction on a large dataset, the standard deviations are generally small across all models. Nonetheless, MS-STGNet consistently achieves higher recall and AUC and lower false alarm rates than STGCN and STGAT under all penetration scenarios, with improvements clearly exceeding the corresponding standard deviations in most cases.

We further conducted paired t-tests across the five runs and found that, for all penetration rates and all reported metrics, the improvements of MS-STGNet over STGCN and STGAT are statistically significant at the 5% level ($p < 0.05$). Section 6.5 has been revised accordingly to highlight both the cross-condition stability and the cross-run robustness of MS-STGNet. Below is the revised version:

Revised:

In Section 6.2 (Page 14 Line 34—35):

To reduce the impact of randomness and evaluate the stability of each method, all models are trained and evaluated five times with different random seeds orders.

In Section 6.5 (Page 16 Line 38—43, Page 17, Page 18 Line 1—43):

To further assess cross-run stability, each entry in Table 4 is reported as the mean ± standard deviation over five independent runs with different random seeds. Statistical tests across the five independent runs show that the improvements of all reported metrics and penetration-rate scenarios are statistically significant at the 5% level ($p < 0.05$).

Traffic conflict prediction remains a significant challenge, particularly in distinguishing between non-conflict and conflict states. Traditional machine learning algorithms, such as SVM and XGBoost, struggle with this task compared to deep learning approaches. For example, under a 30% penetration rate, the recall rates of SVM and XGBoost were 23% and 17.8% lower, respectively, than those of the proposed MS-STGNet. Additionally, their false alarm rates increased by 27.9% and 20.0%, AUC values decreased by 24.3% and 18.9%, and accuracy was reduced by 26.4% and 20.5%. These results emphasize the importance of extracting nonlinear correlations for traffic conflict prediction.

The introduction of deep learning methods significantly improved model performance. CNN and LSTM-CNN outperformed SVM and XGBoost across all metrics, demonstrating the importance of capturing spatial dependencies and temporal correlations in conflict prediction. However, deep learning methods relying on CNNs to capture spatial dependencies face a notable limitation: they cannot model spatial similarities in unconnected grid fields. This highlights the advantage of leveraging graph neural networks (GNNs), such as STGCN and STGAT, to model semantic spatial dependencies, further enhancing performance. For instance, under a 30% penetration rate, STGAT and STGCN improved recall rates by 4.0% and 3.9%, reduced false alarm rates by 2.4% and 3.0%, increased AUC values by 2.9% and 1.5%, and improved accuracy by 4.4% and 3.2%, respectively, compared to LSTM-CNN. These results underscore the advanced capability of utilizing the inherent graph structure of road networks to extract spatial dependencies related to conflict risks. GNNs are particularly well-suited for capturing complex relationships between road segments, integrating heterogeneous road features, and learning network-wide patterns while retaining local details. Comparatively, GAT-based models often outperform GCN models by incorporating predefined adjacency matrices embedded with spatial proximity and contextual similarity, better representing spatial dependencies.

Building on prior advancements in graph-based models, the proposed MS-STGNet model demonstrated robust performance across all penetration rate scenarios. For instance, under a 50% penetration rate, MS-STGNet outperformed the next-best models by 4.9% in recall,

reduced false alarm rates by 3.3%, improved AUC by 5.3%, and increased accuracy by 3.3%. Notably, as shown in Table 4, MS-STGNet achieved a significant reduction in false alarm rates, with improvements of 23.9%, 24.0%, and 23.8% under 50%, 70%, and 90% penetration rates, respectively. This improvement can be attributed to the manifold similarity module, which reduces misjudgments in conflict-prone areas of traffic flow—a point further analyzed in subsequent sections.

Because the task is a binary conflict/non-conflict prediction problem on a large-scale dataset, the standard deviations across runs are generally small for all models. Nevertheless, the reported mean ± standard deviation helps to reveal relative robustness: MS-STGNet maintains consistent advantages over STGCN and STGAT across different penetration rates, and in most cases exhibits comparable or slightly lower variation in key metrics. This indicates that the improvements of MS-STGNet are not due to a single favourable initialization but are reproducible under different random seeds.

Table 4
Performance of Different Models on Datasets.

Penetration rates	Metric	SVM	XGBoost	CNN	LSTM-CNN	STGCN	STGAT	MS-STGNet
10%	Recall	0.531 ±0.049	0.577 ±0.037	0.713 ±0.031	0.726 ±0.024	0.766 ±0.011	0.782 ±0.019	**0.797**↑1.92% ±0.014
	False alarm rate	0.440 ±0.047	0.413 ±0.038	0.206 ±0.029	0.201 ±0.017	0.175 ±0.012	0.165 ±0.009	**0.150**↑9.09% ±0.009
	AUC	0.588 ±0.049	0.632 ±0.039	0.758 ±0.034	0.769 ±0.021	0.790 ±0.020	0.807 ±0.024	**0.824**↑2.11% ±0.016
	Accuracy	0.581 ±0.039	0.652 ±0.055	0.788 ±0.029	0.803 ±0.030	0.830 ±0.014	0.829 ±0.017	**0.855**↑3.01% ±0.011
	G-mean	0.543 ±0.067	0.581 ±0.053	0.745 ±0.037	0.769 ±0.026	0.793 ±0.021	0.803 ±0.016	**0.820**↑2.12% ±0.017
30%	Recall	0.578 ±0.040	0.630 ±0.036	0.742 ±0.022	0.738 ±0.028	0.777 ±0.016	0.778 ±0.013	**0.808**↑3.86% ±0.010
	False alarm rate	0.417 ±0.049	0.338 ±0.054	0.194 ±0.024	0.173 ±0.013	0.143 ±0.013	0.149 ±0.015	**0.138**↑3.50% ±0.013
	AUC	0.596 ±0.047	0.650 ±0.034	0.757 ±0.027	0.781 ±0.025	0.796 ±0.020	0.810 ±0.017	**0.839**↑3.58% ±0.007
	Accuracy	0.592 ±0.065	0.651 ±0.048	0.781 ±0.033	0.789 ±0.018	0.821 ±0.023	0.833 ±0.021	**0.856**↑2.76% ±0.015
	G-mean	0.592 ±0.041	0.644 ±0.046	0.773 ±0.039	0.775 ±0.020	0.815 ±0.014	0.816 ±0.015	**0.831**↑1.84% ±0.012
50%	Recall	0.564 ±0.047	0.593 ±0.054	0.767 ±0.028	0.788 ±0.031	0.803 ±0.023	0.828 ±0.019	**0.877**↑5.92% ±0.009
	False alarm rate	0.417 ±0.047	0.332 ±0.039	0.181 ±0.021	0.165 ±0.019	0.139 ±0.018	0.138 ±0.015	**0.105**↑23.91% ±0.017
	AUC	0.580 ±0.044	0.672 ±0.042	0.790 ±0.036	0.802 ±0.026	0.823 ±0.022	0.833 ±0.024	**0.886**↑6.36% ±0.013
	Accuracy	0.590 ±0.035	0.650 ±0.053	0.794 ±0.032	0.830 ±0.015	0.852 ±0.019	0.857 ±0.016	**0.890**↑3.85% ±0.008
	G-mean	0.563 ±0.050	0.642 ±0.045	0.789 ±0.027	0.813 ±0.030	0.826 ±0.014	0.843 ±0.011	**0.887**↑5.22% ±0.011
70%	Recall	0.571 ±0.038	0.617 ±0.051	0.759 ±0.025	0.749 ±0.028	0.770 ±0.014	0.789 ±0.018	**0.816**↑3.42% ±0.012
	False alarm rate	0.427 ±0.050	0.329 ±0.047	0.168 ±0.030	0.170 ±0.022	0.141 ±0.017	0.125 ±0.012	**0.095**↑24.00% ±0.010
	AUC	0.576 ±0.037	0.673 ±0.050	0.782 ±0.023	0.781 ±0.029	0.816 ±0.020	0.822 ±0.015	**0.860**↑4.62% ±0.015
	Accuracy	0.589 ±0.046	0.668 ±0.057	0.809 ±0.035	0.801 ±0.011	0.830 ±0.013	0.836 ±0.020	**0.898**↑7.42% ±0.007
	G-mean	0.588 ±0.051	0.639 ±0.045	0.802 ±0.028	0.795 ±0.027	0.811 ±0.022	0.828 ±0.009	**0.860**↑3.86% ±0.014
90%	Recall	0.597 ±0.053	0.622 ±0.050	0.782 ±0.034	0.770 ±0.021	0.783 ±0.023	0.809 ±0.010	**0.819**↑1.24% ±0.008
	False alarm rate	0.388 ±0.038	0.352 ±0.041	0.170 ±0.024	0.147 ±0.027	0.130 ±0.016	0.122 ±0.022	**0.093**↑23.77% ±0.011
	AUC	0.595 ±0.040	0.658 ±0.031	0.793 ±0.038	0.786 ±0.012	0.821 ±0.011	0.832 ±0.013	**0.860**↑3.37% ±0.009
	Accuracy	0.591 ±0.063	0.682 ±0.053	0.812 ±0.020	0.835 ±0.024	0.857 ±0.028	0.873 ±0.018	**0.896**↑2.63% ±0.016
	G-mean	0.600 ±0.048	0.635 ±0.035	0.798 ±0.031	0.802 ±0.018	0.822 ±0.019	0.839 ±0.025	**0.863**↑2.86% ±0.013

Comment 4:

As the model will be of interest to traffic-safety practitioners, visualization of the learned manifold-similarity matrices or attention weights could make the results more transparent and explain which spatial-temporal interactions dominate conflict risk.

Response to Comment 4:

We appreciate this constructive suggestion. We agree that visualizing the learned spatial relationships can help practitioners better understand which segment pairs play a dominant role in conflict risk. In response, we have added Appendix A, where we present the learned manifold-similarity matrices for flow, occupancy, and speed. Each matrix is large, so we show the top-left 5×5 block together with the last row and last column, using ellipses to indicate continuation. This tabular layout makes it easier to identify which mainline and ramp segments exhibit strong manifold-based similarity and thus exert greater influence on the learned spatial interactions in MS-STGNet. Below is the revised version:

Revised (Page 26 Line 27—32, Page 27 Line 1—3):

Appendix A. Visualization of the learned manifold-similarity matrices

Appendix A presents the learned manifold-similarity matrices for flow, occupancy, and speed, denoted by $\boldsymbol{Matrices}^{(flow)}$, $\boldsymbol{Matrices}^{(occupancy)}$, *and* $\boldsymbol{Matrices}^{(speed)}$, *respectively. Each matrix is of size* 108×108*; for readability, each matrix lists the top-left* 5×5 *block together with the last row and last column, with ellipses indicating continuation to the full size.*

$$\mathbf{Matrices}^{(\mathrm{flow})} = \begin{bmatrix} 1.000 & 0.277 & 0.268 & 0.274 & 0.745 & \cdots & 0.686 \\ 0.277 & 1.000 & 0.701 & 0.285 & 0.689 & \cdots & 0.279 \\ 0.268 & 0.701 & 1.000 & 0.707 & 0.693 & \cdots & 0.699 \\ 0.274 & 0.285 & 0.707 & 1.000 & 0.688 & \cdots & 0.759 \\ 0.745 & 0.689 & 0.693 & 0.688 & 1.000 & \cdots & 0.696 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.686 & 0.279 & 0.699 & 0.759 & 0.696 & \cdots & 1.000 \end{bmatrix}, \quad \mathbf{Matrices}^{(\mathrm{flow})} \in \mathbb{R}^{108\times108}$$

$$\mathbf{Matrices}^{(\mathrm{occupancy})} = \begin{bmatrix} 1.000 & 0.365 & 0.316 & 0.276 & 0.353 & \cdots & 0.250 \\ 0.365 & 1.000 & 0.367 & 0.327 & 0.302 & \cdots & 0.314 \\ 0.316 & 0.367 & 1.000 & 0.390 & 0.283 & \cdots & 0.358 \\ 0.276 & 0.327 & 0.390 & 1.000 & 0.237 & \cdots & 0.499 \\ 0.353 & 0.302 & 0.283 & 0.237 & 1.000 & \cdots & 0.016 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.250 & 0.314 & 0.358 & 0.499 & 0.016 & \cdots & 1.000 \end{bmatrix}, \quad \mathbf{Matrices}^{(\mathrm{occupancy})} \in \mathbb{R}^{108\times108}$$

$$\mathbf{Matrices}^{(\text{speed})} = \begin{bmatrix} 1.000 & 0.602 & 0.491 & 0.523 & 0.600 & \cdots & 0.396 \\ 0.602 & 1.000 & 0.537 & 0.566 & 0.561 & \cdots & 0.441 \\ 0.491 & 0.537 & 1.000 & 0.503 & 0.450 & \cdots & 0.402 \\ 0.523 & 0.566 & 0.503 & 1.000 & 0.474 & \cdots & 0.512 \\ 0.600 & 0.561 & 0.450 & 0.474 & 1.000 & \cdots & 0.344 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.396 & 0.441 & 0.402 & 0.512 & 0.344 & \cdots & 1.000 \end{bmatrix}, \quad \mathbf{Matrices}^{(\text{speed})} \in \mathbb{R}^{108\times108}$$

Comment 5:

The paper is well written but somewhat lengthy in the literature review. Sections 2.1-2.3 could be tightened without loss of content.

Response to Comment 5:

We thank the reviewer for pointing out that the literature review was somewhat lengthy. We agree that it can be streamlined without losing essential context. Accordingly, we have compressed Sections 2.1–2.3 by 1) merging overlapping descriptions of mixed-traffic simulation and CAV/HDV modelling, 2) summarizing traditional statistical and deep-learning-based safety models more concisely, and 3) grouping early manifold-learning applications into a shorter, synthesized paragraph while retaining detailed discussion of the most relevant recent work. These changes reduce redundancy and improve readability, while still providing sufficient background to motivate MS-STGNet and the proposed manifold-similarity module. Below is the revised version:

Revised:

In Section 2.1 (Page 3 Line 23—43):

Exploring the impact of mixed traffic flow modeling on safety is critical for identifying the key factors required to accurately simulate the driving behaviors of CAVs and HDVs. Existing studies commonly adopt longitudinal car-following models such as Cooperative Adaptive Cruise Control (CACC), Adaptive Cruise Control (ACC) developed by the PATH laboratory (Milanés et al., 2013; Milanés and Shladover, 2014), and the Intelligent Driver Model (IDM) (Treiber et al., 2000) to represent the dynamics of CAVs, autonomous vehicles (AVs), and HDVs in mixed traffic environments (Liu et al., 2018a; Zhou and Zhu, 2020; Yao et al., 2023; Chen et al., 2024b). These models are typically implemented in microscopic traffic simulation tools such as VISSIM, SUMO, and CARLA to evaluate the safety implications of different CAV market penetration rates (MPRs) and traffic demand levels. In general, simulation-based studies report reductions in rear-end and lane-changing conflicts and increases in average travel speeds as CAV/AV penetration increases (Mousavi et al., 2021; Tan et al., 2023). However, several works also highlight that, without advanced V2X communication frameworks and richer behavior modeling, the safety benefits tend to be modest and context-dependent (Tarko, 2021). These findings underscore the importance of integrating realistic vehicle behavior models and communication schemes into mixed-traffic safety assessment frameworks.

A notable gap in these studies is the insufficient distinction between CAVs and HDVs,

particularly in behavioral characteristics such as prolonged reaction times and perceptual uncertainties associated with human drivers, which are often oversimplified in HDV modeling (Gu et al., 2022). While analyses of macroscopic traffic characteristics (e.g., fundamental diagram parameters) may not introduce significant biases, neglecting these distinctions can substantially impact the evaluation of microscopic traffic characteristics, especially those related to safety-critical features (Garg and Bouroche, 2023). In addition, existing conflict or crash prediction models have been rarely tested for their performance in mixed traffic scenarios, leaving a significant gap in understanding their applicability and effectiveness under such complex conditions (Hou et al., 2024a).

In Section 2.2 (Page 3 Line 45—47, Page 4 Line 1—34):

Predicting traffic accidents has long been a critical topic in mobility management research. Early studies predominantly employed traditional statistical methods such as regression models (Caliendo et al., 2007; Bergel-Hayat et al., 2013), Bayesian networks (Martin et al., 2009; Hossain and Muromachi, 2012), and tree-based algorithms (Wang et al., 2010; Lin et al., 2015). These approaches provided initial insights into accident patterns, particularly in small geographical areas, but their ability to capture nonlinear relationships and dynamic dependencies between road segments was limited (Zhang et al., 2014a). Moreover, they often analyzed accident data in isolation, neglecting critical interdependencies between locations, which restricted their applicability to citywide analyses with large datasets (Wang et al., 2021).

With the advent of deep learning, researchers began exploring models that jointly capture spatial and temporal patterns. Convolutional neural networks (CNNs) have been widely used to detect spatial structures (Chen et al., 2018; Hu et al., 2020), while recurrent neural networks (RNNs) and their variants model temporal dependencies (Sameen and Pradhan, 2017; Yuan et al., 2019). Hybrid frameworks such as Long Short-Term Memory (LSTM) networks and ConvLSTM-based architectures further advanced citywide accident prediction by integrating spatial and temporal factors. For example, Ren et al. (2018) used LSTM networks to incorporate temporal influences across multiple locations, and Bao et al. (2019) developed a spatiotemporal convolutional LSTM network (STCL-Net) that effectively captured the spatiotemporal dependencies of urban road networks. However, these grid-based methods often overlooked detailed urban geo-semantic information, such as complex road network semantics and intersection configurations.

To overcome these limitations, graph-based deep learning methods have emerged, leveraging the inherent graph structure of road networks to model spatial relationships. Graph convolutional networks (GCNs) (Zhou et al., 2020; Trirat et al., 2023), graph attention networks (GATs) (Huang et al., 2019; Wang et al., 2023), and spatiotemporal graph neural networks (ST-GNNs) (Yu et al., 2021) have proven effective in integrating spatial and temporal dynamics by representing road segments as nodes and their connections as edges. Several studies have pioneered these advancements. Zhou et al. (2020) introduced the Differential Time-Varying Graph Neural Network (DTGN), integrating spatiotemporal correlations with a data augmentation strategy to address zero inflation in accident data. Yu et al. (2021) proposed a spatiotemporal graph convolutional network featuring a three-layer structure that independently processes the road graph, spatiotemporal data, and embeddings, and tackled

zero inflation by under sampling to balance risky and non-risky segments.

Recent work has further integrated probabilistic frameworks into graph-based models to explicitly account for uncertainty in accident risk. Gao et al. (2024) incorporated Zero-Inflated Tweedie Distributions (ZITD) into an ST-GNN model, parameterizing accident risk with components for mean, variance, and zero inflation to better handle highly imbalanced and long-tailed data. Trirat et al. (2023) proposed a multi-view graph neural network that incorporates both dynamic and static similarity information, providing a more adaptive representation of traffic accidents under dynamic geographical semantics and structural alignment. Their model employs a Huber loss to robustly adapt to zero inflation. Although spatiotemporal GNNs and attention-based adaptive graphs have significantly improved traffic prediction and safety modelling, their applications to real-time conflict prediction in mixed CAV–HDV traffic remain limited, and most adaptive adjacency mechanisms are learned purely from instantaneous node embeddings without an explicit traffic-state prior, which motivates our manifold-similarity-based graph design in the following sections.

In Section 2.3 (Page 5 Line 4—12):

Early studies have applied manifold learning to various traffic-related tasks. For example, Wang et al. (2009) proposed a cooperative traffic state recognition method based on manifold learning that preserves the geometric structure of high-dimensional data, and Lu et al. (2012) introduced a graph embedding algorithm that balances local manifold structures and global discriminative information for traffic sign recognition. Manifold techniques have also been used to identify moving vehicle trajectories and collective behavior patterns. Lee et al. (2012) projected trajectory features onto a 2D manifold and clustered them into a small number of Gaussian components, while Yang and Zhou (2011) combined Local Linear Embedding (LLE) and Principal Component Analysis (PCA) to capture local and global features of traffic parameter data. In addition, Zhang et al. (2014b) employed weighted Euclidean distance based on traffic-parameter similarity to classify traffic states.

Comment 6:

Would it be possible to compare with some (Safe-)RL model?

Response to Comment 6:

We appreciate the reviewer's suggestion to relate our work to (Safe-)RL approaches. We agree that Safe-RL is highly relevant for designing CAV control policies that explicitly account for safety.

In our setting, the task is to estimate, given observed mixed CAV–HDV traffic states, the probability of near-future traffic conflicts at each segment and time step. This is a supervised spatiotemporal prediction problem, for which it is natural to compare against other prediction-based baselines such as STGCN and STGAT. In contrast, (Safe-)RL methods typically learn a control policy (e.g., longitudinal or lane-change decisions, ramp metering, or signal control) by interacting with a simulation environment and optimizing a long-term reward that may include safety-related terms. Their outputs are control actions rather than explicit conflict probabilities, and their performance is evaluated in terms of overall system-level outcomes under the learned

policy.

Because of this fundamental difference in problem formulation and evaluation, a direct, quantitative baseline comparison between MS-STGNet and a Safe-RL controller would not be straightforward or necessarily meaningful: it would require designing a complete CAV control framework, specifying actions and reward functions, and coupling it with a microscopic simulation environment, which goes beyond the scope of the present study. We appreciate the reviewer's suggestion and will consider integrating MS-STGNet into a Safe-RL-based control framework as an important direction for future research.

Comment 7:

Some equations (11-15) may move to an appendix.

Response to Comment 7:

We thank the reviewer for this helpful suggestion. We agree that the detailed mathematical expressions originally given in Eqs. (11)–(15) can be moved to an appendix to improve the readability of the main text. In the revised manuscript, the Gaussian-kernel similarity and AICc-based bandwidth selection from Section 5.3.1 have been relocated to Appendix B.1 (Eqs. (B.1)–(B.2)), and the SVD-based initialization and adaptive adjacency formulation from Section 5.3.2 have been moved to Appendix B.2 (Eqs. (B.3)–(B.5)). The main text now contains concise, high-level descriptions of these steps with explicit references to Appendix B for readers interested in the full derivations. Below is the revised version:

Revised:

In Section 5.3.1 (Page 11 Line 25—30):

By computing the manifold distances between traffic-state vectors across all road segments, we obtain an $n \times n$ geodesic distance matrix. This distance matrix is then converted into a similarity matrix using a Gaussian kernel with bandwidth h. The bandwidth h is automatically selected by minimizing the corrected Akaike Information Criterion (AICc) via a golden-section search. The detailed expressions of the kernel function and the AICc objective are provided in Appendix B (Eqs. (B.1)–(B.2)).

In Section 5.3.2 (Page 12 Line 12—19):

To incorporate potential spatial correlations into our framework, we construct three adaptive graphs by initializing the weights between nodes using similarity matrices. Singular Value Decomposition (SVD) is employed for graph initialization, and the resulting singular components are used to define an initial graph representation. We then introduce learnable left and right transformation matrices, $\boldsymbol{M}_{lt}$ and $\boldsymbol{M}_{rt}$, which operate on the truncated singular vectors and singular values. A nonlinear mapping with ReLU activation and a row-wise softmax is applied to obtain a normalized adaptive adjacency matrix $\widetilde{\boldsymbol{A}^{\star}}$ that balances flexibility and interpretability. The complete mathematical formulation of this SVD-based initialization and adaptive update, including the definitions of $\boldsymbol{M}_{lt}$, $\boldsymbol{M}_{rt}$, and $\widetilde{\boldsymbol{A}^{\star}}$, is given in Appendix B (Eqs. (B.3)–(B.5)).

In Appendix B (Page 27 Line 4—29, Page 28 Line 1—4):

Appendix B. Detailed formulation of manifold-based similarity and adaptive adjacency

B.1. Manifold similarity kernel and bandwidth selection

Given the geodesic distances d_{ij} on the traffic-state manifold, we convert them into a similarity matrix $\mathbf{W}$ *using a Gaussian kernel:*

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{2h^2}\right), \tag{B.1}$$

where d_{ij} represents the manifold distance between traffic states i and j; exp is the exponential function e^x; and h denotes the is kernel bandwidth. The bandwidth h is selected by minimizing the corrected Akaike Information Criterion (AICc) of the resulting model:

$$f(h) = 2k - 2\ln\left(\mathcal{L}(h)\right) + \frac{2k(k+1)}{n-k-1}, \tag{B.2}$$

where n is the sample size, k is the number of free parameters, and $\mathcal{L}(h)$ denotes the likelihood function under bandwidth h.

B.2. SVD-based initialization and adaptive adjacency

To incorporate potential spatial correlations into our framework, we construct three adaptive graphs by initializing the weights between nodes using similarity matrices. Singular Value Decomposition (SVD) is employed for graph initialization (Guo et al., 2015; Zou et al., 2024), and $\mathbf{A}^*$ *can be expressed as the product of three distinct matrices, as follows:*

$$\mathbf{A}^* = \mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*\top} \tag{B.3}$$

where $\mathbf{U}^*$ *and* $\mathbf{V}^*$ *represent orthogonal matrices representing the left and right singular vectors, respectively.* $\mathbf{\Sigma}^*$ *is a diagonal matrix containing singular values. The graph initialized through SVD decomposition provides only a static representation and cannot adapt to the dynamic changes in the data. Therefore, the weight matrix of the adaptive graph,* $\mathbf{A}^\star$*, needs to be optimized through a learnable function:*

$$\mathbf{A}^\star = ReLU(\mathbf{M}_{lt}\mathbf{M}_{rt}) \tag{B.4}$$

where $\mathbf{M}_{lt}$ *and* $\mathbf{M}_{rt}$ *are the core learnable parameter matrices, which play a crucial role in dynamically modeling the weight relationships between nodes in the graph.* $\mathbf{M}_{lt}$ *is the left transformation matrix, designed to encode a linear transformation of the input features or spatial dependency information. It operates as a critical step in updating the representation of node relationships by applying a transformation to the input data, expressed as:* $\mathbf{M}_{lt} = \mathbf{W}_{lt}\left(\widehat{\mathbf{U}}_\star\widehat{\mathbf{\Sigma}}_\star\right)$*. Similarly,* $\mathbf{M}_{rt}$ *is the right transformation matrix, responsible for adjusting or aggregating the information encoded in* $\mathbf{M}_{lt}$*, expressed as:* $\mathbf{M}_{rt} = \mathbf{W}_{rt}\left(\widehat{\mathbf{\Sigma}}_\star\widehat{\mathbf{V}}_\star^\top\right)$*. The ReLU function is applied to introduce nonlinearity and ensure that the weights remain non-negative. Subsequently, the softmax function is used to normalize the weights of each node, ensuring that their sum equals 1. This normalization guarantees a balanced distribution of information during transmission, preventing any single node from dominating the interaction:*

$$\tilde{\mathbf{A}}^\star = \mathbf{I}_N + \mathrm{softmax}\left(\mathrm{ReLU}(\mathbf{M}_{lt}\mathbf{M}_{rt})\right) \tag{B.5}$$

where $\mathbf{I}_N$ *is the identity matrix.*

Comment 8:

English is fluent; minor stylistic tightening would suffice

Response to Comment 8:

We sincerely appreciate the reviewer's positive assessment of the overall English quality. Following your suggestion, we have carefully re-read the entire manuscript and carried out minor stylistic refinements throughout, including polishing sentence structures, improving wording for clarity and conciseness, and harmonizing terminology and notation. We hope that the revised version reads more smoothly and meets the journal's language standards.

Comments from Reviewer #4:

Comment 1:

In Section 4.1, please explain the reasons for selecting a 14-kilometer four-lane highway for simulation.

Response to Comment 1:

We thank the reviewer for this helpful comment. Our simulation setup is closely linked to the calibration of the Enhanced Intelligent Driver Model (EIDM), whose parameters are estimated from the highD dataset of naturalistic trajectories on multilane freeways. For this reason, it is most consistent and reasonable to adopt a freeway scenario rather than urban or suburban roads. The choice of a 14 km four-lane segment reflects a trade-off between realism and computational efficiency: it provides sufficient distance for vehicles to accelerate, cruise, and interact so that stable traffic states and realistic conflicts can develop without being dominated by boundary effects. Several on-ramps and off-ramps are embedded along the segment to mimic real freeway operations and generate complex merging/diverging interactions that are important sources of conflicts in mixed CAV–HDV traffic. We have added a concise explanation of these design choices in Section 4.1 of the revised manuscript. Below is the revised version:

Revised (Page 6 Line 17—23, Page 7 Line 1—2):

This choice is consistent with the calibration of the enhanced intelligent driver model (EIDM), whose parameters are estimated from the highD dataset of naturalistic trajectories on multilane highways. A segment of 14 km provides sufficient distance for vehicles to accelerate, cruise, and interact, so that stable traffic states and realistic conflict events can emerge without being dominated by boundary effects. The main road is segmented into three parts measuring 7,750 m, 3,500 m, and 2,750 m, with speed limits set at 120 km/h, 100 km/h, and 120 km/h, respectively. In addition to the upstream and downstream trunk links, this section includes connections to five on-ramps, featuring a 250-meter-long acceleration lane running parallel, to mimic real freeway operations and to increase the complexity of traffic interactions, thereby generating more representative conflict-prone situations (as shown in Fig.1).

Comment 2:

The paper's research scenario is limited to four-lane highways and does not cover more common urban roads, suburban roads, etc. The authors should consider whether adding pedestrians, non-motorized vehicles, and other elements to the model would significantly impact its performance. Incorporating these elements would enhance the model's robustness in real-world traffic environments. If the authors are unable to include them, this limitation should be explicitly stated in the paper.

Response to Comment 2:

We appreciate this important comment. As clarified in the revised Section 4.1, the present framework is built on an EIDM car-following model calibrated using the highD dataset, which

contains naturalistic trajectories on multilane freeways (See Page 6 Line 16—19). We therefore intentionally choose a freeway segment as the primary simulation environment, since this is the most consistent setting for applying the calibrated model and analysing mixed CAV–HDV interactions at higher speeds.

Our focus in this paper is on segment-level vehicle–vehicle conflict prediction from a macroscopic perspective, rather than on microscopic interactions with pedestrians or non-motorized vehicles. Incorporating these road users would require additional behaviour models and data, and would effectively lead to a different research problem. For example, recent studies on pedestrian–vehicle or motorcycle–pedestrian interactions use dedicated microscopic and reinforcement-learning frameworks tailored to those interactions, rather than freeway segment-level risk prediction (Nasernejad et al., 2021; Lanzaro et al., 2022).

[Lanzaro, G., Sayed, T., & Alsaleh, R. (2022). Can motorcyclist behavior in traffic conflicts be modeled? A deep reinforcement learning approach for motorcycle-pedestrian interactions. *Transportmetrica B: transport dynamics*, *10*(1), 396-420.\

Nasernejad, P., Sayed, T., & Alsaleh, R. (2021). Modeling pedestrian behavior in pedestrian-vehicle near misses: A continuous Gaussian Process Inverse Reinforcement Learning (GP-IRL) approach. *Accident Analysis & Prevention*, 161, 106355.]

At the same time, we agree that additional contextual factors (e.g., weather, pavement friction, POIs) are relevant for real-world conflict risk. These variables are more aligned with our macroscopic setting but cannot be fully and reliably represented in the current simulation. We therefore focus on core traffic-flow variables—time of day, flow, speed, and occupancy—as inputs, which we regard as a simple yet effective choice for freeway mixed-traffic scenarios.

Following the reviewer's suggestion, we have made this limitation explicit in the Conclusion. The revised text now 1) states that the model is calibrated and evaluated only on a four-lane freeway with motorized traffic, and that direct transferability to urban or suburban networks with pedestrians and non-motorized vehicles is limited, and 2) refines the practical implications to focus on freeway and urban-expressway safety management, which better matches the simulation setting.Below is the revised version:

Revised (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalised intersections, pedestrians, and non-motorised vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the

framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring scalable pretraining and training strategies on larger and more diverse networks, including freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

Comment 3:

MS-STGNet does not explicitly specify which traffic features contribute most significantly to conflict prediction. I suggest the authors add this information to Section 6.

Response to Comment 3:

We thank the reviewer for this insightful comment. All models in our study, including MS-STGNet, take as inputs macroscopic traffic features (flow, mean speed, occupancy) together with the CAV penetration rate, which are processed jointly within the spatiotemporal graph network. We note that Section 6.5 already compares model performance across five CAV penetration scenarios and highlights penetration rate as a key scenario variable in mixed CAV–HDV traffic (See Page 18 Line 21—24); in the revised version we make this role more explicit and connect it to the feature-oriented discussion in Section 6.10.

Section 6.10 and Appendix C have been expanded to analyse how traffic volume and resulting speed-dispersion patterns affect conflict risk and model behaviour. Using 500 hours of simulation with low/medium/high volume levels, we show that higher volumes produce more pronounced speed oscillations along the segment and substantially higher empirical conflict rates. The supplementary trajectory plots (Figures C1–C3) highlight that HDVs and heavy vehicles generate larger and more frequent speed fluctuations, while increasing CAV penetration smooths these perturbations. Combined with the segment-level risk profiles in Fig. 9, these results indicate that CAV penetration rate, traffic volume, and speed separation are among the most influential traffic features for conflict prediction, and that MS-STGNet aligns its predicted risk with these structures more reliably than STGCN and STGAT.

We also note that deep spatiotemporal graph networks do not natively provide scalar feature-importance scores as in tree-based models; a full attribution analysis (e.g., SHAP or GNN-specific explainability) is thus left for future work. Below is the revised version:

Revised:

In Section 6.10 (Page 23 Line 28—29, Page 24 Line 1—17):

Beyond penetration rates, we also examined how traffic volume and the resulting speed dispersion patterns affect conflict risk and model behavior. In the simulation, different representative demand levels were considered over a total of 500 hours, covering low-, medium-, and high-volume conditions. The supplementary trajectory plots in Appendix C (Figures C.1–C.3) show that as traffic volume increases, pronounced speed oscillations emerge along the segment and become more frequent and severe. This indicates that, even under mixed CAV–HDV conditions, higher demand intensifies vehicle interactions and amplifies the likelihood of conflicts, which supports our use of traffic state variations as predictors of conflict occurrence. A closer inspection of these trajectories further highlights the role of different vehicle classes and CAV penetration as key traffic features. The green and blue trajectories representing HDVs exhibit larger amplitude and higher-frequency speed fluctuations than the red trajectories representing CAVs, reflecting more aggressive driving behavior and delayed responses in the human-driven fleet. Heavy vehicles (trucks) introduce additional instability due to their limited acceleration and deceleration capabilities and larger size, which force surrounding vehicles to adjust their speeds more frequently and create pronounced perturbation zones. As CAV penetration increases, these unstable zones shrink and the gaps between high-speed and low-speed vehicle clusters are gradually bridged by heterogeneous CACC queues, leading to smoother trajectories and reduced speed dispersion. Combined with the segment-level risk profiles in Fig.9, these observations indicate that CAV penetration rate, traffic volume, and the resulting speed separation patterns are among the most influential traffic features for conflict prediction in the proposed framework: MS-STGNet is particularly effective at aligning its predicted risk with these underlying speed dispersion structures, while STGCN and STGAT tend to generate spurious conflict probabilities in disturbance zones.

In Appendix C (Page 28 Line 7—13, Page 29):

Appendix C. Supplementary vehicle position–speed trajectories

In this appendix, we provide additional vehicle position–speed trajectory plots for three representative demand levels, corresponding to low-, medium-, and high-volume conditions. For each traffic volume, the trajectories are shown separately for the pre-merging, merging, and post-merging segments, with different colors indicating HDVs, CAVs, and heavy vehicles (trucks). These plots illustrate how increasing traffic volume and changes in vehicle composition lead to more pronounced speed oscillations and perturbation zones, complementing the case study around Fig. 9 in the main text and supporting the discussion in Section 6.10 on the impact of traffic volume, CAV penetration, and speed separation on conflict risk.

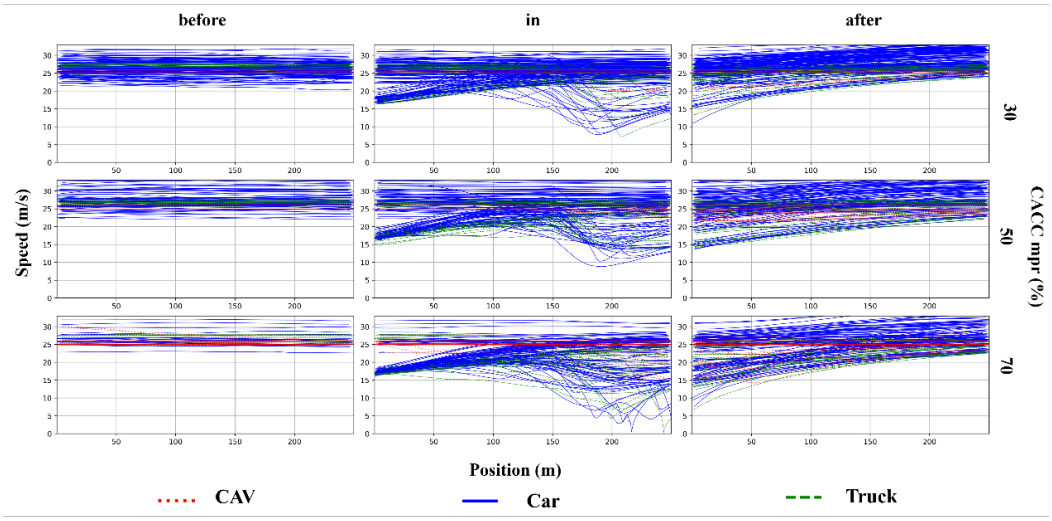

Fig. C1. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 3000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

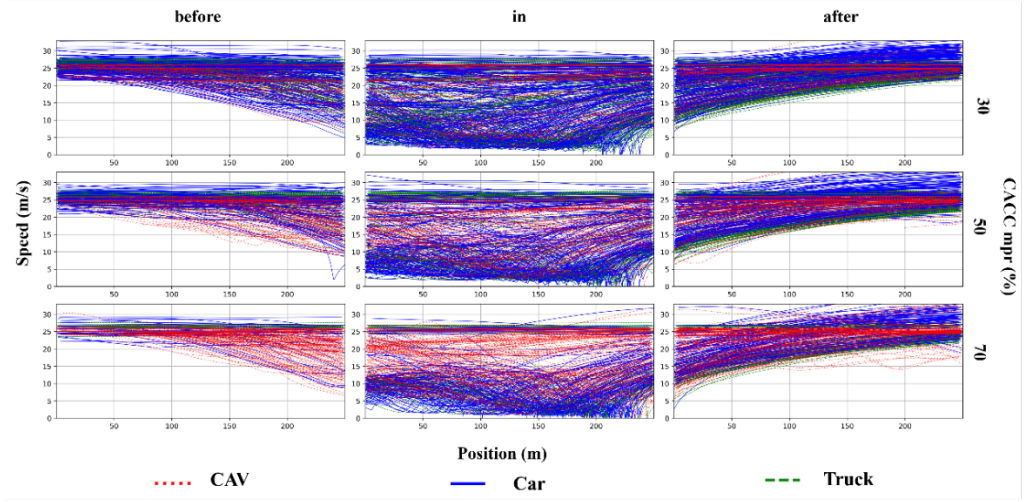

Fig. C2. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 6000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

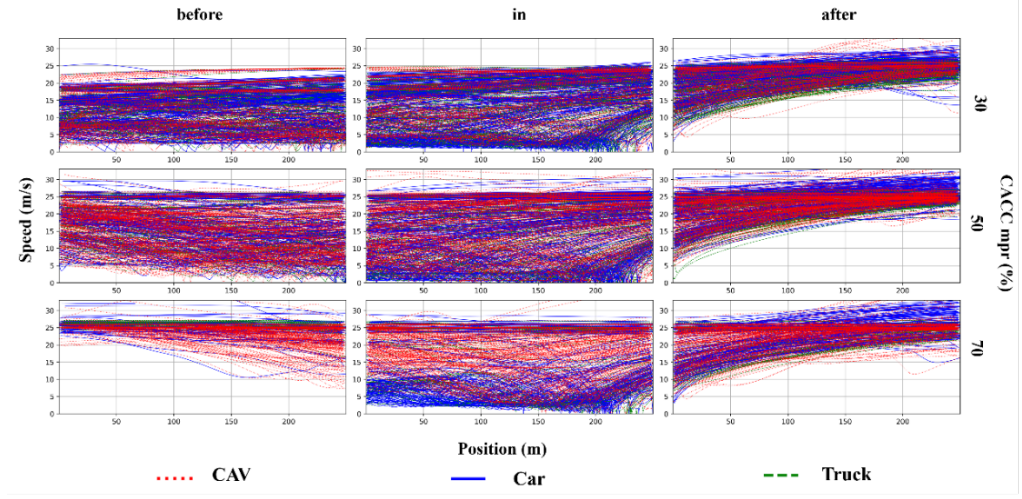

Fig. C3. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 9000

vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

Comment 4:

The finding of insufficient dialogue with existing research and it is suggested that the authors strengthen the discussion of the current study's results with the existing literature after obtaining the results in Section 6.5.

Response to Comment 4:

We appreciate this constructive suggestion and agree that the empirical findings should be more clearly related to existing work. In the revised manuscript, we have strengthened the end of Section 6.5 in two respects: 1) We explicitly note that our results are consistent with previous studies showing that spatiotemporal graph models such as STGCN and STGAT generally outperform traditional machine-learning and sequence models for traffic prediction tasks. 2) We clarify that MS-STGNet extends these STGNN approaches by embedding a manifold-based state-similarity prior into the adaptive graph learning process. We also connect our findings to recent manifold-learning studies in traffic flow and safety analysis, emphasizing that our framework complements this line of work by demonstrating how manifold-informed similarity can be integrated into a deep spatiotemporal graph network for mixed CAV–HDV conflict prediction. We hope this enhanced discussion in Section 6.5 addresses the reviewer's concern about insufficient dialogue with existing research. Below is the revised version:

Revised (Page 18 Line 31—43):

These empirical results also clarify how MS-STGNet differs in practice from STGAT-type adaptive graph models. Although both approaches employ graph-based representations, STGAT relies on feature-driven attention to construct adjacency at each time step, which can be sensitive to local fluctuations in highly imbalanced conflict datasets. By contrast, MS-STGNet constrains the adaptive graph updates within a manifold-similarity prior derived from historical traffic states. As the market penetration of CAVs increases and pronounced speed separation emerges, this manifold-informed prior helps the model avoid spuriously high conflict probabilities in non-conflict regions, leading to consistently lower false alarm rates and more stable performance across all penetration scenarios. In this sense, our findings are consistent with previous studies showing that graph-based spatiotemporal models such as STGCN and STGAT outperform traditional machine-learning and sequence models in traffic prediction tasks, while further extending them by explicitly incorporating a manifold-based state similarity prior into the adaptive graph learning process. At the same time, our results complement recent manifold-learning approaches for traffic state analysis by demonstrating that manifold-informed similarity can be embedded into deep spatiotemporal graph networks to improve conflict prediction in mixed CAV–HDV freeway traffic.

Comment 5:

The authors' current research focuses solely on dichotomous predictions of conflict presence or absence, without addressing graded predictions of conflict severity (e.g., minor scrape risk versus severe collision risk). This is also a limitation of the study.

Response to Comment 5:

We are grateful for this thoughtful comment and agree that modelling graded conflict severity would further enhance practical relevance. In the current study, we formulate the problem as binary conflict/non-conflict prediction because our simulation provides reliable labels for conflict occurrence, but not well-validated categorical labels for different severity levels.

We would like to clarify, however, that the output of MS-STGNet is a continuous risk score in [0,1] obtained via a sigmoid activation. The threshold of 0.5 is used only for computing classification metrics (See Page 16 Line 2—5); the raw probabilities are used directly to construct the segment-level risk profiles in Section 6.10 and can already be interpreted as a continuous measure of conflict risk intensity.

Nevertheless, we fully acknowledge that we do not explicitly design or evaluate a multi-level severity model in this work. In the revised Conclusion, we have (i) added this point to the list of limitations, stating that the framework currently focuses on binary prediction, and (ii) explicitly highlighted, in the future work paragraph, the extension from binary conflict detection to graded or ordinal conflict severity prediction by combining the continuous risk scores with appropriate severity labels (e.g., minor, moderate, severe). Below is the revised version:

Revised (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalised intersections, pedestrians, and non-motorized vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring scalable pretraining and training strategies on larger and more diverse networks, including

freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

Comments from Reviewer #5:

Comment 1:

The evaluation is conducted using simulated traffic datasets generated with SUMO. Although the authors claim to have calibrated their simulation models (EIDM and CACC) using real-world datasets, the conflict data analyzed are synthetic. It is recommended to validate the proposed MS-STGNet model with a real-world traffic dataset (such as highD, NGSIM, etc.) where traffic conflicts can be extracted or inferred.

Response to Comment 1:

We sincerely appreciate this important comment. At the early stage of this study, our initial intention was indeed to develop and validate MS-STGNet directly on real-world mixed CAV–HDV data. However, after a thorough review of existing datasets, we found that currently available data sources cannot simultaneously meet the two core requirements of our problem: 1) truly mixed CAV–HDV traffic, and 2) long, spatially continuous freeway segments with macroscopic measurements (flow, speed, occupancy) suitable for segment-level conflict prediction over continuous time series.

On the one hand, classical trajectory datasets such as NGSIM and highD contain only human-driven vehicles and therefore do not match our target mixed-traffic scenario. Nevertheless, to ensure that our simulation is not based on purely theoretical assumptions, we calibrated the human-driven car-following model directly on highD freeway trajectories, so that HDV behaviour in the simulation reflects empirically observed acceleration, deceleration, and headway patterns rather than arbitrary parameter choices.

On the other hand, recent autonomous-vehicle datasets such as the Lyft Level 5 AV Dataset (Houston et al., 2021), nuScenes (Caesar et al., 2020), and the Waymo Open Dataset (Sun et al., 2020) do provide mixed traffic with AVs/CAVs, but their structure is not well suited to our macroscopic conflict-prediction task. As summarized in Table 1 of Hu et al. (2022), these AV datasets are organized into short trajectory segments: Waymo comprises 1,000 segments with a temporal resolution of 0.1 s and a typical segment length of 20 s; Lyft Level 5 contains 366 segments at 0.2 s resolution and 25–45 s duration; and nuScenes includes 1,000 segments with 0.5 s resolution and 20 s duration. These segments are collected from the viewpoint of individual AVs and are neither spatially contiguous along a single freeway facility nor temporally continuous over long periods. Hu et al. (2022) explicitly note that substantial preprocessing and reconstruction are required even to obtain usable car-following trajectories from these segment-based recordings, and that the resulting data remain fragmented in space and time for macroscopic analyses.

Table 1. Overview of three AV trajectory dataset.

Dataset	Number of segments	Resolution (s)	Length of each segment (s)
Waymo	1000	0.1	20
Lyft	366	0.2	25–45
nuScenes	1000	0.5	20

[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).

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., ... & Ondruska, P. (2021, October). One thousand and one hours: Self-driving motion prediction dataset. In Conference on Robot Learning (pp. 409-418). PMLR.

Hu, X., Zheng, Z., Chen, D., Zhang, X., & Sun, J. (2022). Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research. *Transportation Research Part C: Emerging Technologies*, *134*, 103490.

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).]

Similarly, Zhang et al. (2025) show that such AV trajectory datasets are particularly suitable for microscopic behaviour analysis and data-driven stochastic fundamental-diagram modelling, but they are inherently sparse in space and time and therefore not directly aligned with macroscopic, segment-level modelling of traffic states over extended freeway sections. In our study, however, the prediction target is whether a given freeway segment will experience a conflict within a continuous time series, based on macroscopic indicators (flow, speed, occupancy) monitored along a 14 km stretch. This requires long, contiguous observations along one facility, which current AV datasets do not provide.

[Zhang, X., Yang, K., Sun, J., & Sun, J. (2025). Stochastic fundamental diagram modeling of mixed traffic flow: A data-driven approach. *Transportation Research Part C: Emerging Technologies*, *179*, 105279.]

These limitations are consistent with the broader challenges identified in recent reviews of machine-learning-based crash prediction. For example, Ali et al. (2024) point out that empirical safety studies for mixed CAV–HDV environments are still scarce, that most real-time crash and conflict prediction models are developed for conventional freeway or urban networks without CAVs, and that many studies necessarily rely on simulation or indirectly inferred data due to the lack of suitable field datasets. Against this background, we adopted a hybrid strategy that combines empirical calibration with large-scale simulation.

[Ali, Y., Hussain, F., & Haque, M. M. (2024). Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention*, *194*, 107378.]

Given these data limitations, we adopted a hybrid "real-trajectory calibration + simulation-based testing" strategy. At the microscopic level, we construct mixed CAV–HDV traffic with multiple penetration rates (10%, 30%, 50%, 70%, and 90%) and generate trajectories at 0.2s resolution. Using these trajectories, we compute widely used surrogate safety measures TTC, DRAC, and DDR—to identify both longitudinal and lateral conflicts. At the macroscopic level, we aggregate the simulated data along a 14 km four-lane freeway with ramps and extract continuous time series of flow, speed, and occupancy for each segment, paired with the conflict/non-conflict labels derived from TTC/DRAC/DDR. This design allows us to study conflict prediction as a segment-level time-series classification problem under a wide range of demand levels and CAV penetration rates, while keeping driver behavior anchored in real freeway observations and defining conflicts via physically interpretable criteria.

We fully acknowledge that this hybrid validation cannot replace large-scale field testing on real mixed CAV–HDV networks. In the revised Conclusion, we now explicitly state that 1)

the current evaluation is conducted in a calibrated freeway simulation, 2) the direct transferability to urban or suburban networks is therefore limited, and 3) the lack of suitable real-world mixed-traffic datasets is a key limitation of the present work. We also clarify that, once continuous macroscopic observations of mixed CAV–HDV traffic over long freeway segments become available, we plan to retrain and evaluate MS-STGNet on those data and to systematically compare its performance with other state-of-the-art safety models in real-world environments. Below is the revised version:

Revised (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalized intersections, pedestrians, and non-motorized vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring scalable pretraining and training strategies on larger and more diverse networks, including freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

Comment 2:

It is advisable to test the model in different traffic conditions (low, medium, and high congestion), different road geometries (e.g., merging ramps), or even an urban environment.

Response to Comment 2:

We thank the reviewer for this constructive suggestion. In the revised manuscript, we clarify more explicitly how varying traffic conditions and geometries are already incorporated, and we delimit the current scope to freeway environments.

Section 6.1 explicitly states that the simulation covers 500 hours with hourly demand randomly sampled between 2,500 and 10,000 veh/h and stratified into low-, medium-, and high-volume ranges (See Page 14 Line 10—16). Section 6.10 and Appendix C have been expanded with additional analyses: the supplementary position–speed trajectories (Figures C.1–C.3) show that increasing volume leads to stronger speed oscillations and higher conflict occurrence, supporting our use of traffic-state variations as predictors.

Section 4.1 has been clarified to emphasize that the four-lane freeway includes multiple on-ramps with 250 m acceleration lanes (See Page 6 Line 22—23, Page 7 Line 1—2). The case study in Section 6.10 (Fig. 9) is conducted on one such merging segment, explicitly distinguishing pre-merging, merging, and post-merging subsegments.

We agree that extending the framework to urban or suburban networks would further enhance its scope. However, the present study is built on an EIDM model calibrated with highD freeway trajectories, making a freeway corridor the most consistent evaluation setting (See Page 6 Line 16—19). Extending MS-STGNet to urban networks would require additional behaviour models and data sources and is therefore left as future work. This limitation and planned extension are now explicitly stated in the Conclusion. Below is the revised version:

Revised:

In Section 6.10 (Page 23 Line 28—29, Page 24 Line 1—17):

Beyond penetration rates, we also examined how traffic volume and the resulting speed dispersion patterns affect conflict risk and model behavior. In the simulation, different representative demand levels were considered over a total of 500 hours, covering low-, medium-, and high-volume conditions. The supplementary trajectory plots in Appendix C (Figures C.1–C.3) show that as traffic volume increases, pronounced speed oscillations emerge along the segment and become more frequent and severe. This indicates that, even under mixed CAV–HDV conditions, higher demand intensifies vehicle interactions and amplifies the likelihood of conflicts, which supports our use of traffic state variations as predictors of conflict occurrence. A closer inspection of these trajectories further highlights the role of different vehicle classes and CAV penetration as key traffic features. The green and blue trajectories representing HDVs exhibit larger amplitude and higher-frequency speed fluctuations than the red trajectories representing CAVs, reflecting more aggressive driving behavior and delayed responses in the human-driven fleet. Heavy vehicles (trucks) introduce additional instability due to their limited acceleration and deceleration capabilities and larger size, which force surrounding vehicles to adjust their speeds more frequently and create pronounced perturbation zones. As CAV penetration increases, these unstable zones shrink and the gaps between high-speed and low-speed vehicle clusters are gradually bridged by heterogeneous CACC queues, leading to smoother trajectories and reduced speed dispersion. Combined with the segment-level risk profiles in Fig.9, these observations indicate that CAV penetration rate, traffic volume, and the resulting speed separation patterns are among the most influential traffic features for conflict prediction in the proposed framework: MS-STGNet is particularly effective at aligning its predicted risk with these underlying speed dispersion structures, while

STGCN and STGAT tend to generate spurious conflict probabilities in disturbance zones.

In Appendix C (Page 28 Line 7—13, Page 29):

Appendix C. Supplementary vehicle position–speed trajectories

In this appendix, we provide additional vehicle position–speed trajectory plots for three representative demand levels, corresponding to low-, medium-, and high-volume conditions. For each traffic volume, the trajectories are shown separately for the pre-merging, merging, and post-merging segments, with different colors indicating HDVs, CAVs, and heavy vehicles (trucks). These plots illustrate how increasing traffic volume and changes in vehicle composition lead to more pronounced speed oscillations and perturbation zones, complementing the case study around Fig. 9 in the main text and supporting the discussion in Section 6.10 on the impact of traffic volume, CAV penetration, and speed separation on conflict risk.

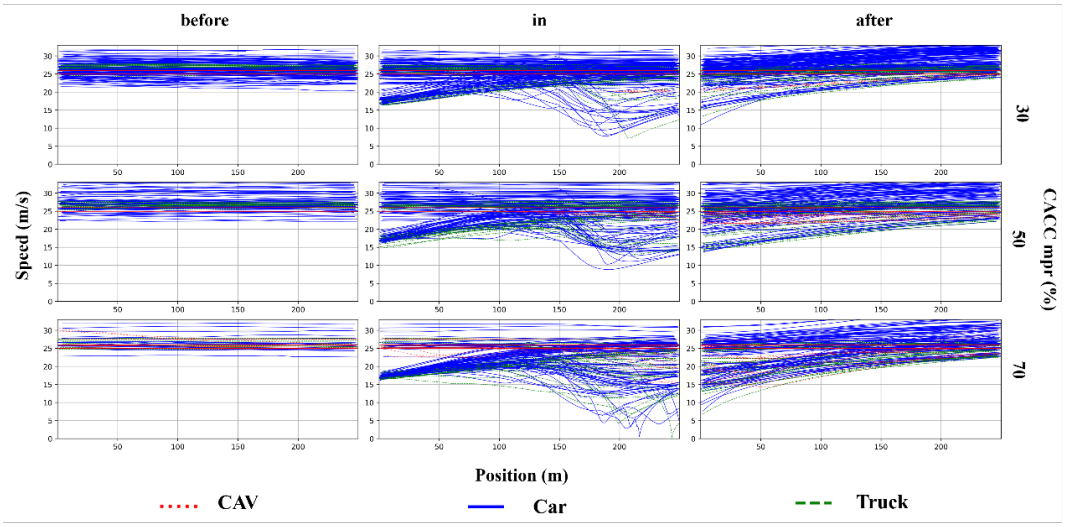

Fig. C1. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 3000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

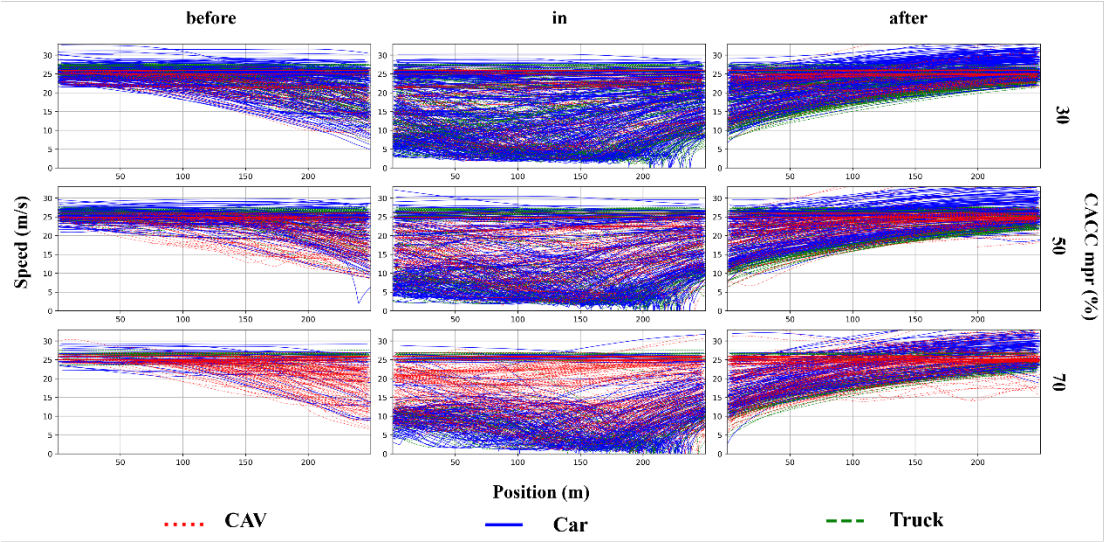

Fig. C2. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 6000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

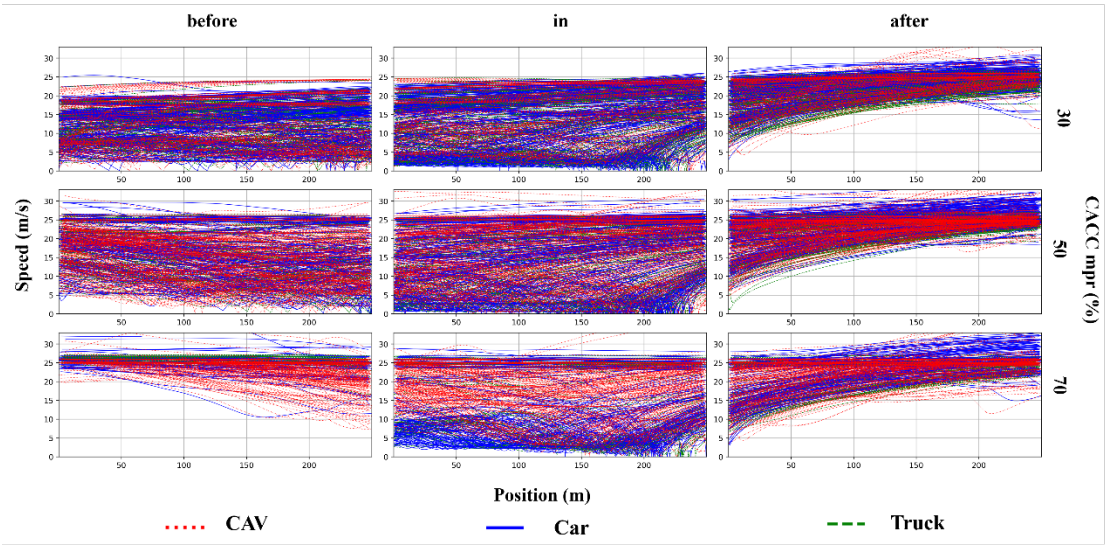

Fig. C3. Vehicle position-speed trajectories at different penetration rates with a traffic volume of 9000 vehicles/hour. (before) pre-merging segment. (in) merging segment. (after) post-merging segment.

In Section 7 (Page 25 Line 25, Page 26 Line 1—21):

The proposed framework has several practical implications. It can be embedded as a safety prediction component in CAV cloud management systems for freeway corridors and urban expressways, integrated into freeway traffic management centers and ramp control or variable speed limit systems to support mixed CAV–HDV operations, and used within regional expressway operation platforms to provide real-time conflict or crash risk warnings at bottlenecks and merging/diverging areas, thereby enhancing the safety management and visualization of freeway networks. The limitations of this study are summarized as follows: 1) The model is calibrated and evaluated in a microscopic simulation of a four-lane freeway segment with motorized traffic only. Although the simulation is grounded in highD trajectory data, we do not yet validate MS-STGNet on large-scale field observations of mixed CAV–HDV traffic, and the direct transferability of the results to urban or suburban road networks with signalised intersections, pedestrians, and non-motorised vehicles is therefore limited. 2) The current experiments focus on a single 14 km corridor with specific demand patterns; additional facilities and more diverse demand scenarios would further test the generalizability of the framework. 3) The predefined manifold similarity matrix remains static over time, preventing the model from capturing previously unseen traffic state transitions unless it is retrained. 4) The proposed framework currently focuses on binary conflict/non-conflict prediction. Although the sigmoid activation in the output layer produces continuous risk scores in the [0,1] range, we do not explicitly model or evaluate graded levels of conflict severity (e.g., minor versus severe conflicts). Moving forward, future works contain: 1) Collecting or leveraging emerging mixed CAV–HDV field datasets with continuous monitoring, so as to retrain and validate MS-STGNet under real-world conditions and assess its scalability. 2) Developing online or adaptive manifold-learning strategies to update similarity matrices in real time. 3) Exploring scalable pretraining and training strategies on larger and more diverse networks, including

freeway corridors and urban expressways with additional contextual variables such as weather conditions, pavement friction, and points of interest (POIs). 4) Extending MS-STGNet from binary conflict detection to graded or ordinal conflict severity prediction by combining continuous risk scores with appropriate severity labels.

Comment 3:

The proposed architecture uses multiple components (convolutional residual, manifold similarity, TCN, and adaptive fusion). Processing these components could incur significant computational overhead for training and real-time inference. Could the authors provide an analysis on the computational cost of this framework?

Response to Comment 3:

We appreciate this important comment and agree that computational cost is crucial for practical deployment. In response, we have added a dedicated computational-cost comparison in Section 6.6 (Table 5), reporting for all deep models under five penetration rates: 1) GPU-MUT (peak GPU memory during training), 2) GPU-MUI (peak GPU memory during inference), and 3) the number of trainable parameters. For SVM and XGBoost, GPU-based indicators are omitted because they run on CPU and have negligible memory usage compared to deep models.

The results show that STGCN consistently has the largest parameter count and GPU memory footprint, with STGAT slightly smaller but still heavier than CNN and LSTM-CNN. By contrast, MS-STGNet uses fewer parameters than both graph-based baselines and reduces peak GPU memory by roughly 10–15% in training and 15–25% in inference across penetration rates, while still incorporating the manifold-similarity module and adaptive fusion. Compared with CNN/LSTM-CNN, MS-STGNet incurs moderately higher memory usage due to graph operations but remains in the same order of magnitude and does not introduce prohibitive overhead.

We also note that the manifold similarity matrices are computed once offline from historical data; training and inference operate on a fixed sparse manifold graph using standard spatiotemporal graph and temporal convolutions. Given this design and the hardware-agnostic indicators in Table 5, we believe that MS-STGNet achieves a reasonable balance between accuracy (Table 4) and efficiency, making it suitable for practical mixed CAV–HDV conflict prediction. We do not report wall-clock times since they depend heavily on specific hardware/software environments and are difficult to reproduce.

Revised (Page 18 Line 44—52, Page 19 Line 1—16):

6.6. Computation cost

In real-world deployment, predictive accuracy is the primary requirement for traffic safety applications, while the hardware cost of the deployed model constitutes a secondary but still crucial consideration for practical implementation. To highlight the computational overhead of different approaches, Table 5 reports three indicators under five CAV penetration-rate scenarios: GPU-MUT (peak GPU memory usage during training), GPU-MUI (peak GPU memory usage during inference), and the number of trainable parameters. For the classical machine-learning baselines (SVM and XGBoost), GPU-based indicators are omitted ("–")

because they are trained and executed on CPU and their memory footprint is negligible compared with deep models in our setting.

Several observations can be made from Table 5. First, among the deep learning baselines, STGCN consistently has the largest parameter count and highest GPU memory usage, with STGAT slightly smaller but still noticeably heavier than CNN and LSTM-CNN. For example, at a 50% penetration rate, STGCN and STGAT require 479,816 and 426,572 parameters, respectively, and their GPU-MUT values reach 4,497 MiB and 4,681 MiB. By contrast, the proposed MS-STGNet uses fewer parameters than both graph-based baselines (395,428 at 50% penetration) and reduces peak GPU memory by roughly 10–15% in training (e.g., 4,059 MiB versus 4,497–4,681 MiB) and 15–25% in inference (e.g., 2,710 MiB versus 3,216–3,587 MiB), while still incorporating a manifold-similarity module and adaptive fusion. Compared with CNN and LSTM-CNN, MS-STGNet understandably incurs moderately higher GPU memory usage due to the additional graph operations, but remains in the same order of magnitude and does not introduce prohibitive overhead.

Table 5
The computational performance of different models on dataset.

Penetration rates	Metric	SVM	XGBoost	CNN	LSTM-CNN	STGCN	STGAT	MS-STGNet
10%	GPU-MUT	—	—	4,333MiB	4,443MiB	5,574MiB	5,802MiB	**5,031MiB**
	GPU-MUI	—	—	2,283MiB	2,799MiB	4,446MiB	3,986MiB	**3,359MiB**
	Parameters	—	—	298,742	346,251	594,758	528,759	**490,154**
30%	GPU-MUT	—	—	4,419MiB	4,530MiB	5,684MiB	5,917MiB	**5,130MiB**
	GPU-MUI	—	—	2,328MiB	2,854MiB	4,534MiB	4,065MiB	**3,425MiB**
	Parameters	—	—	304,621	353,064	606,462	539,164	**499,800**
50%	GPU-MUT	—	—	3,496MiB	3,584MiB	4,497MiB	4,681MiB	**4,059MiB**
	GPU-MUI	—	—	1,842MiB	2,258MiB	3,587MiB	3,216MiB	**2,710MiB**
	Parameters	—	—	241,008	279,335	479,816	426,572	**395,428**
70%	GPU-MUT	—	—	3,085MiB	3,162MiB	3,968MiB	4,130MiB	**3,581MiB**
	GPU-MUI	—	—	1,625MiB	1,992MiB	3,165MiB	2,838MiB	**2,391MiB**
	Parameters	—	—	212,656	246,474	423,370	376,390	**348,909**
90%	GPU-MUT	—	—	2,983MiB	3,058MiB	3,837MiB	3,994MiB	**3,463MiB**
	GPU-MUI	—	—	1,572MiB	1,927MiB	3,061MiB	2,744MiB	**2,312MiB**
	Parameters	—	—	205,636	238,338	409,394	363,965	**337,392**

Overall, these results indicate that MS-STGNet achieves superior predictive performance (as shown in Table 4) with a computational cost that is only modestly higher than conventional CNN-based models and clearly lower than that of STGCN and STGAT. This suggests that the proposed architecture strikes a reasonable balance between accuracy and efficiency, making it suitable for deployment in practical mixed CAV–HDV conflict prediction systems. We do not report wall-clock training or inference time, as such measurements are highly dependent on specific hardware, software environments, and background system load; instead, we focus on parameter counts and GPU memory usage, which provide hardware-agnostic indicators of computational complexity.

Comments from Reviewer #6:

Comments:

This manuscript presents a deep learning-based approach to stress detection in software developers using EEG signal analysis. The subject is timely and relevant, given the growing concern about mental well-being in tech-heavy occupations. The study leverages deep learning algorithms and biomedical signal processing, aligning well with current trends in AI-driven healthcare and human-centered computing.

Strengths

The paper addresses an important interdisciplinary problem combining mental health, software engineering, and AI.

Use of EEG signals provides a physiological and objective measure of stress.

The authors apply deep learning models, which are state-of-the-art for classification tasks, and present reasonable performance metrics.

Weaknesses and Suggestions

Experimental Setup: The study would benefit from a clearer description of dataset size, sampling procedures, and participant demographics.

Comparative Analysis: No strong benchmarking with baseline models or traditional ML classifiers (e.g., SVM, Random Forest). This limits the understanding of the added value of using DL.

Model Explainability: There is limited discussion of how interpretable the model's decisions are. In biomedical contexts, explainability is vital.

Writing and Structure: While generally well-organized, there are some grammatical errors and vague phrasing in the results and discussion sections that require revision.

Discussion of Limitations: The manuscript lacks a critical reflection on the generalizability of the model and the potential for bias in data collection.

Response to Reviewer #6:

We sincerely thank the reviewer for the time and effort devoted to evaluating our submission. However, we respectfully note that this specific set of comments does not appear to refer to our manuscript. Our paper focuses on traffic conflict prediction in mixed CAV–HDV freeway environments using a manifold similarity-based spatiotemporal graph neural network (MS-STGNet). The manuscript does not involve EEG data, stress detection in software developers, biomedical signal analysis, or AI-driven healthcare applications. Given this clear mismatch in topic, data, and methodology, we are unfortunately unable to provide a meaningful point-by-point response to the detailed remarks in this particular review comment, as they do not correspond to the content of our work.