Griffin: Aerial-Ground Cooperative Detection and Tracking Dataset and Benchmark

Jiahao Wang[1], Xiangyu Cao[1], Jiaru Zhong[2], Yuner Zhang[3], Zeyu Han[1], Haibao Yu[4], Chuang Zhang[1], Lei He[1], Shaobing Xu[1] *, Jianqiang Wang[1] *

[1]Tsinghua University
[2]The Hong Kong Polytechnic University
[3]University of Pennsylvania
[4]The University of Hong Kong

Abstract

While cooperative perception can overcome the limitations of single-vehicle systems, the practical implementation of vehicle-to-vehicle and vehicle-to-infrastructure systems is often impeded by significant economic barriers. Aerial-ground cooperation (AGC), which pairs ground vehicles with drones, presents a more economically viable and rapidly deployable alternative. However, this emerging field has been held back by a critical lack of high-quality public datasets and benchmarks. To bridge this gap, we present *Griffin*, a comprehensive AGC 3D perception dataset, featuring over 250 dynamic scenes (37k+ frames). It incorporates varied drone altitudes (20-60m), diverse weather conditions, realistic drone dynamics via CARLA-AirSim co-simulation, and critical occlusion-aware 3D annotations. Accompanying the dataset is a unified benchmarking framework for cooperative detection and tracking, with protocols to evaluate communication efficiency, altitude adaptability, and robustness to communication latency, data loss and localization noise. By experiments through different cooperative paradigms, we demonstrate the effectiveness and limitations of current methods and provide crucial insights for future research.

Code — https://github.com/wang-jh18-SVM/Griffin

1 Introduction

While significant progress has been made in autonomous driving technologies, current single-vehicle systems still struggle with fundamental challenges of severe occlusions and limited field-of-view in complex environments. To address these limitations, cooperative perception strategies, including vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I), have emerged, offering substantial improvements in perception capabilities. Nevertheless, their practical implementation often requires significant infrastructure investments and widespread adoption of connected vehicles, which can present substantial economic barriers. In contrast, vehicle-to-drone or so called, aerial-ground cooperative (AGC) systems, which integrate unmanned aerial vehicles (UAVs) with ground vehicles, provide a unique alternative. They leverage on-demand drone deployment and an unobstructed bird's-eye view, offering a flexible, economical solution for critical environments, including smart cities, emergency response, and security patrols.

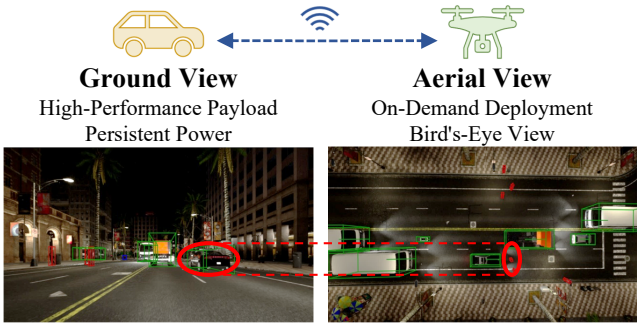


Figure 1: Motivation for aerial-ground cooperative perception. AGC provides a flexible option for cooperative perception by leveraging on-demand deployment and a unique bird's-eye view. In this example, the aerial view reveals pedestrians (red circle) occluded from the vehicle.

Despite the promising potential of AGC perception systems, relevant progress is hindered by a critical shortage of high-quality, representative benchmarks. This gap stems from the inherent complexities of aerial-ground interaction, which pose significant challenges for both real-world data collection and high-fidelity simulation. In real-world scenarios, the dynamic perspective between aerial and ground sensors complicates annotation and calibration. Unlike V2V or V2I systems, where sensors generally move on a horizontal plane, UAVs introduce complex motion with continuous changes in altitude, pitch, and roll, which disrupt precise cross-view correspondence. Consequently, many efforts have turned to simulation to bypass these challenges. However, existing simulation-based datasets still fail to replicate real-world complexities. For instance, many datasets oversimplify the scene with ideal localization and communication (Ye, Sunderraman, and Ji 2024; Tian et al. 2024; Wang et al. 2024b, 2025; Gao et al. 2025), or employ simplistic drone models with fixed orientation (Wang et al. 2024b,

*Corresponding authors: Jianqiang Wang and Shaobing Xu (wjqlws@tsinghua.edu.cn, shaobxu@tsinghua.edu.cn)

Mode	Dataset	Source	BBox Type	Cameras per Agent	Sequential Tracking	Occlusion Aware	Realistic Noise	Frames (k)	Altitude (m)
Veh-Veh	OPV2V(ICRA 2022)	Joint Sim	3D	Multiple	✓	✓	×	11	–
	V2V4Real(CVPR 2023)	Real	3D	Multiple	✓	–	–	310	–
Veh-Inf	DAIR-V2X(CVPR 2022)	Real	3D	Single	×	–	–	22	20-25
	V2X-Seq(CVPR 2023)	Real	3D	Single	✓	–	–	15	20-25
Air-Air	CoPerception-UAVs(NIPS 2022)	Joint Sim	3D	Multiple	✓	×	✓	5.2	40,60,80
	UAV3D(NIPS 2024)	Joint Sim	3D	Multiple	✓	×	×	20	60
	AeroCollab3D(TGRS 2024)	Joint Sim	3D	Single	✓	×	×	3.2	50
	Air-Co-Pred(NIPS 2024)	Sim	3D	Single	✓	✓	×	8.0	50
Veh-Air	CoPeD(RAL 2024)	Real	2D	Single	×	–	–	203	2-10
	V2U-COO†(TGRS 2025)	Sim	3D	Single	×	×	✓	9.3	70, 80
	AGC-VUC†(Preprint)	Real	3D	Multiple	×	–	–	20	10-15
	AirV2X-Perception(Preprint)	Joint Sim	3D	Multiple	✓	×	✓	121.1	60-105
	Griffin (Ours)	**Joint Sim**	**3D**	**Multiple**	✓	✓	✓	**37.7**	**20-60**

Table 1: Comparison of representative cooperative perception datasets. Our dataset, *Griffin*, is highlighted as the only one in the vehicle-aerial domain to support occlusion-aware annotations and realistic noise simulation. In the Source column, 'Joint Sim' denotes co-simulation of CARLA and AirSim / SUMO; 'Sim' uses only CARLA. 'Occlusion Aware' shows if annotation visibility is considered for simulated data. 'Realistic Noise' indicates supports for simulating communication interference and localization errors. 'Altitude' represents the height of infrastructure or drone sensors. †Attributes are derived from the publications as the datasets are not released yet. Symbols: ✓ (Supported), × (Not Supported), – (Not Applicable/Specified).

2025) or constant altitudes (Ye, Sunderraman, and Ji 2024; Tian et al. 2024; Wang et al. 2024b, 2025). Furthermore, many of them (Hu et al. 2022; Ye, Sunderraman, and Ji 2024; Tian et al. 2024; Wang et al. 2025; Gao et al. 2025) lack occlusion-aware annotations, resulting in labels for invisible objects. Overcoming these shortcomings is crucial for developing robust AGC perception systems applicable to real-world environments.

Inspired by the Griffin, a mythical creature that unites the lion's terrestrial strength and the eagle's aerial dominance, we aim to harness the combined power of aerial and ground perspectives to overcome these challenges and enhance collaborative ability. To this end, we present the following contributions for aerial-ground cooperative 3D perception:

- **The Griffin Dataset**: We release *Griffin*, an aerial-ground cooperative 3D perception dataset. It encompasses over 250 dynamic scenes (37K frames, 340K images) from CARLA-AirSim co-simulation, with instance-aware occlusion quantification, varying cruising altitudes, and realistic simulation of drone dynamics under various conditions.
- **A Comprehensive Benchmark**: We present a benchmarking framework for evaluating aerial-ground cooperative 3D object detection and tracking. It includes implementations of classic baselines and provides a suite of metrics to evaluate accuracy, communication cost, and robustness under varying communication interference and localization errors.

2 Related Work

Cooperative perception research has been significantly propelled by datasets spanning various communication modes, as detailed in Table 1. For V2V and V2I scenarios, benchmarks range from comprehensive simulations (Xu et al. 2022b; Li et al. 2022a) to real-world collections (Yu et al. 2022; Xu et al. 2023; Yu et al. 2023; Hao et al. 2024). In contrast, the cooperative perception datasets featuring aerial perspectives from UAVs remain limited. Existing works, such as UAV3D (Ye, Sunderraman, and Ji 2024), AeroCollab3D (Tian et al. 2024), and Air-Co-Pred (Wang et al. 2024b) primarily focus on Air-Air cooperation and are often constrained by idealized communication and localization or simplified drone dynamics, such as fixed altitudes.

For the more challenging AGC scenarios, pioneering datasets have emerged, though with notable limitations. CoPeD (Zhou et al. 2024), while large-scale, targets low-altitude robot scenarios and employs unrefined, automatically generated annotations. V2U-COO (Wang et al. 2024a, 2025) relies on predefined UAV poses that lack realistic motion dynamics. More recent efforts have also sought to advance the field but leave critical gaps. AGC-Drive (Hou et al. 2025) lacks tracking IDs for temporal analysis and is constrained to low-altitude flights, while AirV2X (Gao et al. 2025) omits crucial occlusion-aware annotations.

To address these limitations, our work introduces *Griffin*, a comprehensive solution designed to advance the development of deployable aerial-ground perception systems. Our dataset bridges the aforementioned gaps by providing realistic multi-agent dynamics through co-simulation, complete with occlusion-aware 3D annotations and tracking IDs across diverse altitude settings. Furthermore, we introduce a robust benchmarking framework engineered to bridge the sim-to-real gap, which allows a systematic robustness evaluation against controllable, real-world imperfections.

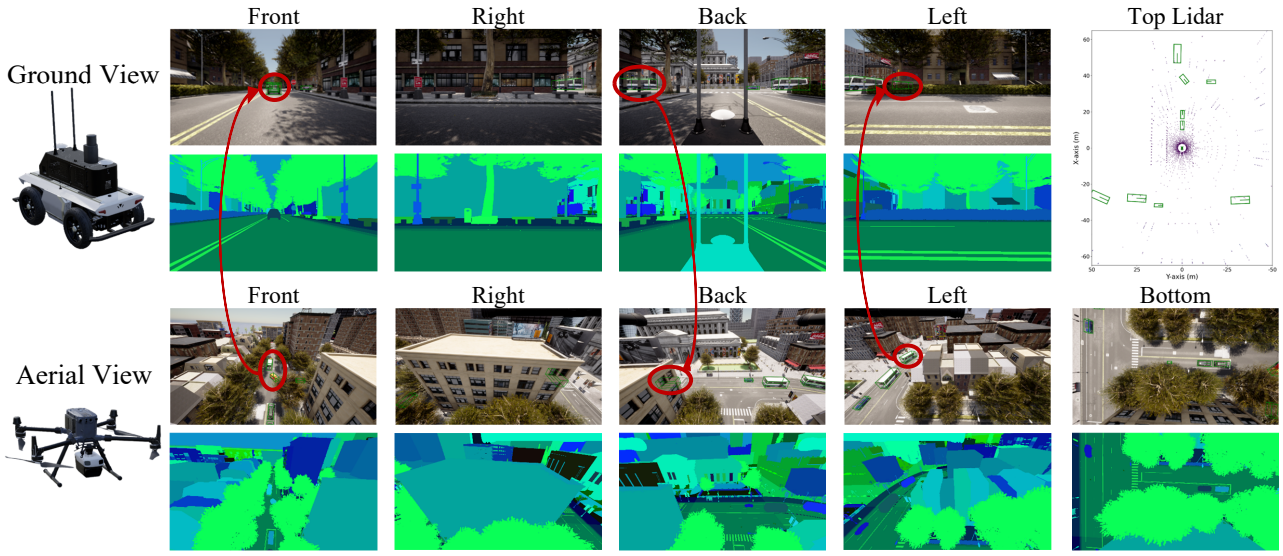


Figure 2: An example from *Griffin* with visualized annotation. The ground vehicle platform is equipped with four cameras and one LiDAR, while the aerial drone platform has five cameras. We also provide instance segmentation ground truth, as shown in the lower row. Bounding boxes represent annotations from cooperative perspectives, indicating that one agent should be able to 'see' certain occluded objects after communication with the other. We use red circles and arrows to highlight those cases.

3 Data Setup

3.1 Data Collection

To generate synchronized multi-agent scenarios, we employ a co-simulation framework which leverages CARLA (Dosovitskiy et al. 2017) for its rich environmental maps, dynamic traffic flows, and high-fidelity sensors, complemented by AirSim's (Shah et al. 2018) realistic, physics-based modeling of drone dynamics.

Sensor Configuration. The sensor suites for both ground and aerial platforms are carefully designed to balance perceptual capabilities with platform-specific constraints. As illustrated in Figure 2, the ground platform features a multimodal sensor suite with four wide field-of-view (FoV) RGB cameras (108.8°, 1920×1080 resolution) and an 80-beam LiDAR operating at 10 Hz with a vertical FoV from -25° to 15°. In contrast, strict size, weight, and power (SWaP) constraints for the aerial platform necessitate a vision-centric configuration without LiDAR. Consequently, the aerial platform employs five downward-oriented cameras with sensor specifications matching those of the ground vehicle.

Scene Diversity. To ensure robustness and generalizability, the dataset was collected across a wide range of simulated environments and conditions. Data were captured from four representative CARLA maps—two urban (Town03, Town10HD) and two suburban (Town06, Town07)—with varying actor densities and vehicle speeds. As detailed in Figure 3, scenes feature diverse weather, including different times of day (noon, sunset, night), clarity levels (clear, rainy, foggy), and wind speeds (0–9 m/s).

Based on UAV cruising altitude, the dataset is organized into four categories. The *Griffin-Random* features the widest altitude range, from 20 to 60 meters above the vehicle. In contrast, *Griffin-25m*, *Griffin-40m*, and *Griffin-55m* focus on specific altitude bands at 25 ± 2m, 40 ± 2m, and 55 ± 2m, respectively. In total, the dataset comprises 255 scene clips, each lasting approximately 15 seconds, totaling over 37.7k samples, 339.3k images, and 914.8k 3D annotations.

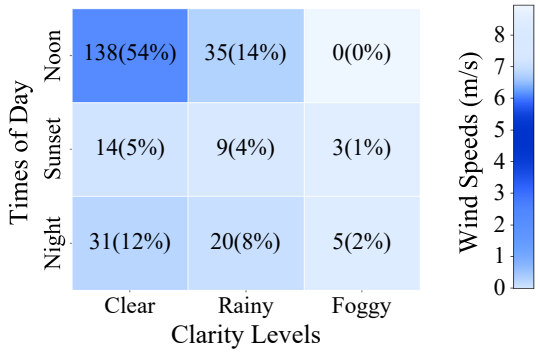
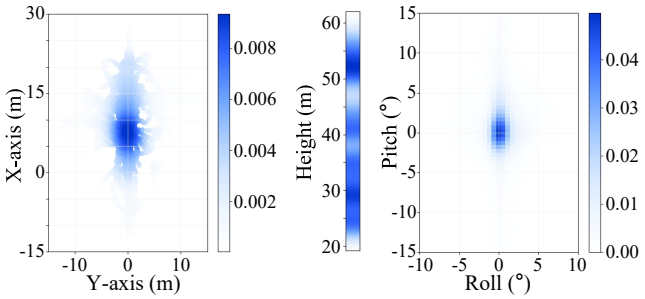


Figure 3: Weather distribution of scene clips. The dataset encompasses a variety of weather and lighting conditions. Following real-world patterns, certain combinations, such as fog at noon, are intentionally rare or absent.

Furthermore, we designed multiple collaboration modes by varying the horizontal and vertical distances between the ground vehicle and UAV, creating diverse relative positioning patterns. As shown for *Griffin-Random* in Figure 4, the drone is typically positioned several meters ahead of the vehicle, serving as a forward scout. The realism of our simulation is further reflected in the drone's orientation angles. The distributions for roll and pitch angles show variance around zero rather than being sharply peaked, which demonstrates the drone's continuous micro-adjustments to reach acceleration targets and maintain stability against simulated wind.



(a) Horizontal Position (Relative to Ego Vehicle) (b) Vertical Position (c) Orientation Angle (Pitch/Roll)

Figure 4: UAV pose distribution of *Griffin-Random*.

3.2 Data Post-Processing

Spatio-Temporal Alignment. The *Griffin* dataset involves four coordinate system categories: world, ego, sensor, and simulator, as detailed in Table 2. Our spatial alignment pipeline converts all simulator-native 3D annotations into a unified right-handed coordinate system. Dual output formats are available that support both the ego-centric KITTI benchmark (Geiger et al. 2013) and the global-reference NuScenes benchmark (Caesar et al. 2020).

To ensure temporal consistency, CARLA's synchronous mode was activated during data recording, which guarantees that all data captured share a precise timestamp.

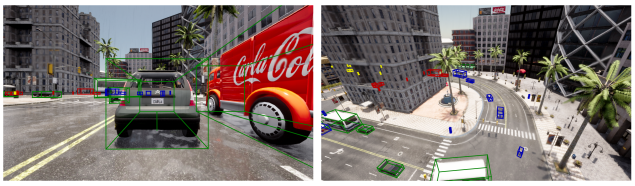
To bridge the sim-to-real gap, we also provide a code interface to inject user-specified real-world imperfections during training and inference, including localization errors, communication latency, and packet loss, enabling a full evaluation of model robustness.

Occlusion-Aware Annotation. *Griffin* provides high-quality, frame-by-frame 3D annotations for six object categories: pedestrian, car, truck, bus, motorcycle, and bicycle. Each annotation includes a category label, a persistent ID, a calculated visibility rate, and a 9-DoF bounding box defined by x, y, z, length, width, height, roll, pitch, and yaw.

To address the critical challenge of judging occlusions, we developed a visibility quantification method that leverages CARLA's instance segmentation ground truth interface. During data collection, RGB and segmentation images are recorded simultaneously using identical sensor configurations to ensure perfect spatio-temporal alignment. In post-processing, we sample points within each ground-truth bounding box and project them onto the segmentation image. Visibility rates for each agent are then calculated by comparing semantic categories and instance IDs of the sampled pixels against the corresponding target's. Targets with low visibility from a single agent's perspective are filtered out to ensure annotation precision. For the final cooperative perception ground truth, targets visible to either agent are retained, as illustrated with green and blue boxes in Figure 5.

Name	Category	Type	Origin
World	Geodetic	ENU (R)	Fixed reference point
Ego	Drone	FLU (R)	Drone center
	Vehicle	FLU(R)	Vehicle center
Sensor	Camera	RDF (R)	Camera optical center
	LiDAR	FLU (R)	LiDAR center
Simulator	CARLA	ESU (L)	Fixed reference point
	AirSim	NED (R)	Fixed reference point

Table 2: Different coordinate systems. Axis direction: ENU (East-North-Up), FLU (Forward-Left-Up), RDF (Right-Down-Forward), RD (Right-Down), ESU (East-South-Up), NED (North-East-Down). Handedness: R (Right-handed), L (Left-handed)



(a) Vehicle Front View (b) Drone Front View

Figure 5: Unfiltered annotations regardless of visibility. Boxes are color-coded by their visibility to show the necessity of our occlusion-aware filtering. Green boxes represent targets visible to the vehicle, while blue boxes are those made visible by the drone's complementary view; both are retained as the ground truth for cooperative perception. In contrast, many existing datasets only filter by distance (yellow boxes) and neglect heavily occluded targets (red boxes).

4 Tasks, Metrics, and Baselines

The *Griffin* dataset is designed to support a variety of cooperative perception tasks, including but not limited to 3D object detection, tracking, motion prediction, and semantic segmentation. In this paper, we narrow our focus to two fundamental visual tasks: 3D Object Detection and Tracking.

4.1 AGC 3D Object Detection Task

This task requires the detection of 3D objects within a predefined region of interest R_g around the ego-vehicle using cooperative data from ground (g) and aerial (a) platforms. For a perception timestamp T_g, inputs consist of image sequences $\{C_g(t_g) \mid t_g \leq T_g\}$ and $\{C_a(t_a) \mid t_a \leq T_a\}$, where $C(\cdot)$ denotes the capture function, along with the relative agent pose $M_{a_{T_a} \to g_{T_g}}$. The desired output is a set of detections, each containing a 3D bounding box, a semantic label, and a confidence score. The corresponding ground truth, GT_{detect}, is formulated by uniting the annotated objects visible to either agent and then filtering for those within the region of interest:

$$GT_{\text{detect}} = \{o \mid o \in GT_g \cup GT_a \text{ and } \text{center}(o) \in R_g\}$$

where each object o is defined by its 3D bounding box and label, and $\text{center}(o)$ is the box's geometric center.

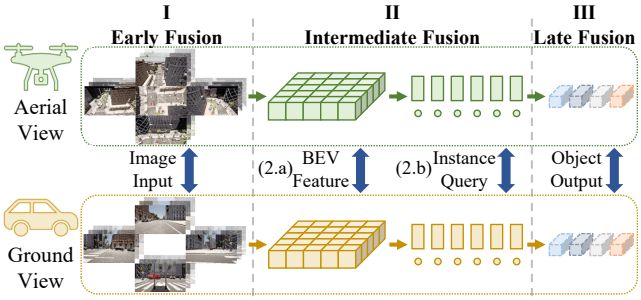


Figure 6: Overview of the cooperative fusion paradigms

4.2 AGC 3D Object Tracking Task

The tracking task extends detection by requiring a model to maintain a unique and persistent identity for each object over time. Our framework supports two standard tracking paradigms: a *joint tracking* approach that uses the same raw sensor inputs as the detection task, and a *tracking-by-detection* approach that uses pre-computed detections as input. In both cases, the output for each object must augment its 3D bounding box and label with a persistent tracking ID. The corresponding ground truth, GT_{track}, is structured accordingly to include these IDs.

4.3 Evaluation Metrics

To ensure a comprehensive and standardized evaluation, we adopt established metrics from the NuScenes benchmark (Caesar et al. 2020) for 3D object detection and tracking, including Average Precision (AP) to assess detection quality and Average Multi-Object Tracking Accuracy (AMOTA) for tracking performance. Beyond perception accuracy, we also assess the communication efficiency by quantifying the data transmitted in Bytes per second (BPS). This allows for a direct analysis of each method's trade-offs by comparing its performance improvement over a no-fusion baseline against its required communication bandwidth.

4.4 Baseline Framework

We implement and evaluate a series of baseline methods to establish performance references for AGC 3D object detection and tracking. As shown in Figure 6, existing cooperative perception methods can be categorized by fusion stages. We provide implementations for all four major paradigms, utilizing a unified BEVFormer backbone (Li et al. 2022b) for fair comparison.

For intermediate fusion, which balances performance and communication costs, we evaluate four methods across two levels. At the BEV level, we implement V2X-ViT (Xu et al. 2022a) and Where2comm (Hu et al. 2022), which reduce bandwidth through feature compression and confidence-based selection, respectively. At the instance level, we employ UniV2X (Yu et al. 2025) and CoopTrack (Zhong et al. 2025). They exchange sparse object queries and perform cross-agent association using a rule-based Hungarian algorithm (Kuhn 1955) and a learnable module, respectively. These, along with standard Early and Late Fusion implementations, form a comprehensive framework to objectively compare different fusion strategies for AGC perception.

5 Experiments

The detection and tracking performance of different fusion methods across the Griffin datasets are presented in Table 3, which details the AP and AMOTA scores alongside the associated communication costs. The experiments reveal distinct performance characteristics for each fusion strategy.

5.1 Performance Overview

As expected, Early Fusion consistently establishes the upper performance bound, achieving the highest AP and AMOTA scores across all test conditions, albeit at an exceptionally high communication cost of 3.11×10^8 BPS (311 MB/s). Conversely, Late Fusion represents the lower performance bound, offering only marginal gains and, in some cases, underperforming the No Fusion baseline. However, due to its minimal communication cost of just 1.56×10^3 BPS, its modest improvements yield the highest gain-per-byte efficiency among all methods when a positive gain is achieved.

Intermediate fusion methods provide a balance between these extremes. Among the methods with a communication cost under 1 MB/s, CoopTrack stands out, providing significant performance gains over the baseline at a moderate communication cost of 1.17×10^5 BPS. V2X-ViT also demonstrates strong performance but requires a higher communication bandwidth due to its reliance on dense, scene-level BEV feature transmission. Surprisingly, Where2comm and UniV2X yield unsatisfactory performance gains. We attribute this to the inherent sparsity of targets from the drone's aerial perspective. For methods like Where2comm, which use positive detections to generate spatial confidence maps for compressing BEV features, this sparsity can lead to the loss of valuable information. Similarly, for sparse query-based methods like UniV2X, it can result in an insufficient number of matched positive samples during training, thereby limiting the model's learning capacity. In contrast, although CoopTrack is also based on sparse object queries, its learnable instance association module leverages ground-truth matching relationships, enabling more effective cross-domain alignment and association.

5.2 Generalization to Flight Altitude Changes

The results in Table 3 also demonstrate that the performance of cooperative perception methods is sensitive to UAV altitude, varying significantly across different flight heights.

First, all cooperative methods achieve their highest performance gains on the low-altitude *Griffin-25m* dataset. As altitude increases in the *Griffin-40m* and *Griffin-55m* datasets, performance degrades across all methods. This trend underscores a fundamental challenge in aerial perception: as altitude increases, the reduced apparent scale of ground targets makes them progressively more difficult to detect. Interestingly, on the *Griffin-Random* dataset, which features UAV altitudes ranging from 20m to 60m, most fusion methods perform worse than the No Fusion baseline. This suggests that the varying altitudes, and the resulting inconsistencies

Method	Griffin-25m		Griffin-40m		Griffin-55m		Griffin-Random		Comm. Cost (BPS)	Comp. Eff. (FPS)
	AP	AMOTA	AP	AMOTA	AP	AMOTA	AP	AMOTA		
No Fusion	0.375	0.365	0.341	0.363	0.335	0.359	0.459	0.481	0	8.10
Early Fusion	**0.607** (+0.232)	**0.670** (+0.305)	**0.503** (+0.162)	**0.555** (+0.192)	**0.483** (+0.148)	**0.522** (+0.163)	**0.583** (+0.124)	**0.649** (+0.168)	3.11×10^8	5.17
V2X-ViT (ECCV 2022)	0.465 (+0.090)	0.508 (+0.143)	0.410 (+0.069)	0.502 (+0.139)	0.350 (+0.015)	0.379 (+0.020)	0.400 (-0.059)	0.423 (-0.058)	8.00×10^5	7.56
Where2Comm (NIPS 2022)	0.396 (+0.021)	0.406 (+0.041)	0.345 (+0.004)	0.413 (+0.050)	0.317 (-0.018)	0.353 (-0.006)	0.406 (-0.053)	0.451 (-0.030)	3.30×10^5	7.60
CoopTrack (ICCV 2025)	0.479 (+0.104)	0.488 (+0.123)	0.396 (+0.055)	0.446 (+0.083)	0.364 (+0.029)	0.402 (+0.043)	0.468 (+0.009)	0.490 (+0.009)	1.17×10^5	6.23
UniV2X (AAAI 2025)	0.419 (+0.044)	0.456 (+0.091)	0.348 (+0.007)	0.401 (+0.038)	0.323 (-0.012)	0.349 (-0.010)	0.402 (-0.057)	0.443 (-0.038)	5.58×10^4	7.06
Late Fusion	0.378 (+0.003)	0.377 (+0.012)	0.335 (-0.006)	0.391 (+0.028)	0.306 (-0.029)	0.332 (-0.027)	0.375 (-0.084)	0.400 (-0.081)	1.56×10^3	6.83

Table 3: Model performance, communication cost, and computational efficiency. The Frames Per Second (FPS) values are measured on a single NVIDIA 3090 GPU. Parenthesized values show absolute gain over the No Fusion baseline. Bold values denote the best overall performance. Underlined values indicate the highest gain-per-byte efficiency.

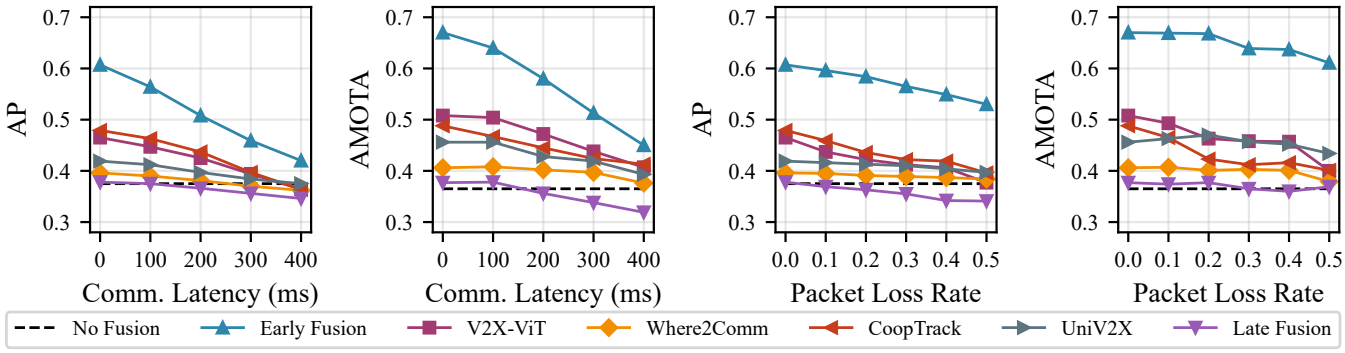


Figure 7: Robustness to communication interference.

in target scale and position, pose a critical challenge to the network's generalization capabilities, overriding the benefits from a lower average altitude or larger dataset size compared to *Griffin-55m*. These findings highlight the necessity for more adaptive fusion mechanisms that can accommodate dynamic UAV perspectives.

Furthermore, different fusion strategies exhibit distinct robustness profiles to altitude variations. Instance-level methods prove more resilient, as CoopTrack is the only intermediate fusion method that maintains a performance advantage over the No Fusion baseline on the challenging *Griffin-Random* dataset. This gap can be attributed to their differing approaches to geometric transformation. BEV-level fusion demands a rigid alignment of dense, geometrically-sensitive feature grids, a process vulnerable to the scale and perspective distortions caused by varying altitudes. In contrast, instance-level methods flexibly transform the sparse 3D reference points associated with each object query while preserving their semantic features. This decoupling of geometry and semantics makes the fusion process inherently more resilient to spatial inconsistencies.

5.3 Communication Robustness

To simulate real-world communication challenges, we evaluate the robustness of different fusion strategies against network imperfections, specifically communication latency and packet loss. While existing studies (Kutila et al. 2021) report typical latencies of 20-50 ms and packet loss rates (PLR) around 10% under standard conditions, our evaluation intentionally explores a more demanding range—up to 400 ms latency and 50% PLR. This aggressive testing is designed to identify the failure points of each fusion strategy, which is particularly critical for aerial-ground scenarios where drones may encounter more severe signal interference and intermittent connectivity than their ground-based counterparts.

The results, illustrated in Figure 7, reveal inherent trade-offs between fusion paradigms and communication reliability. Early Fusion, which transmits large volumes of raw im-

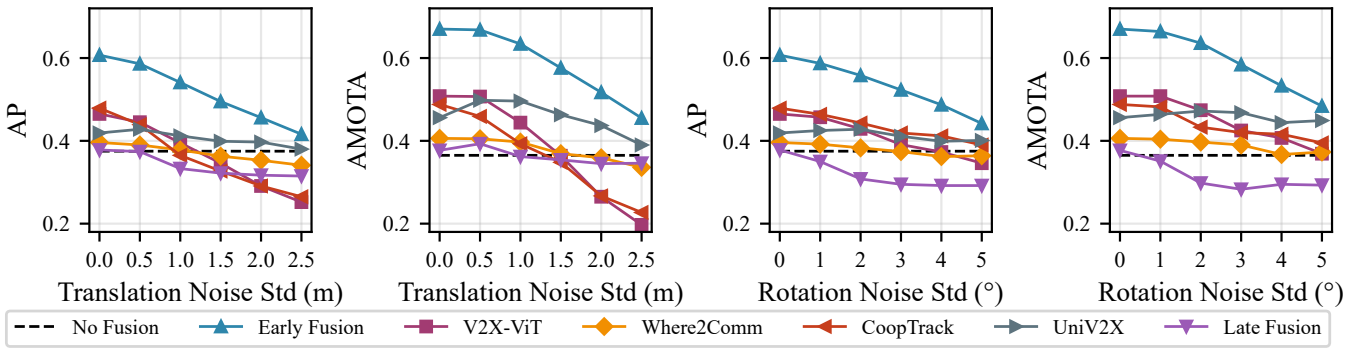


Figure 8: Robustness to localization error.

age data, is highly susceptible to latency, with its AP score dropping by over 30% at 400 ms. In contrast, intermediate fusion methods exhibit better resilience. While their detection performance surpasses the No Fusion baseline at latencies up to 200 ms, their tracking capabilities are even more robust, maintaining a consistent advantage across all tested latency levels up to 400 ms. As for packet loss, its influence appears less detrimental than that of latency. Even under severe conditions, such as a 50% packet loss rate, most fusion methods maintain a performance advantage over the No Fusion baseline. We hypothesize that this is because dropped packets lead to a loss of information, reducing potential performance gains, but do not introduce erroneous data that could generate additional false-positive signals.

5.4 Localization Robustness

We investigate the impact of localization errors by introducing noise into the UAV's transformation matrix, separately analyzing translation and rotation errors. While existing benchmarks (Xu et al. 2022a; Hu et al. 2022; Wang et al. 2025) often assess noise within narrow ranges (e.g., 0.6m or 1.0°), we follow CBM (Song et al. 2024) and adopt a more challenging evaluation framework, injecting Gaussian noise with standard deviations (std) of up to 2.5 meters for translation and 5 degrees for rotation. This rigorous testing is crucial for understanding model reliability with severe GPS inaccuracies or calibration drifts.

As shown in Figure 8, the performance of most cooperative methods is highly sensitive to both translation and rotation errors. Among the intermediate fusion methods, V2X-ViT and CoopTrack suffer the most significant degradation, with their performance dropping below the No Fusion baseline when the translation error exceeds 1.5m. In contrast, UniV2X maintains a clear advantage over the No Fusion baseline across all tested error levels, and Where2comm also exhibits a more graceful degradation. We attribute their superior robustness to their selective use of transmitted data, employing instance-level filtering or spatial confidence maps to down-weight unreliable signals.

6 Discussion and Conclusion

This paper introduces the Griffin framework, a novel dataset and benchmark designed to accelerate research in aerial-ground cooperative 3D perception. The experiments demonstrate the significant potential of this paradigm. In favorable conditions, cooperative methods achieve substantial performance gains over single-agent baselines by resolving occlusions and expanding the effective field-of-view.

However, this work also underscores that the full potential of AGC is yet to be realized, as the performance gains of current fusion methods are fragile and highly dependent on idealized conditions. Two primary challenges are identified: a strong sensitivity to the drone's flight altitude and a significant vulnerability to real-world imperfections like communication interference and localization errors. Further analysis provides critical insights into these issues. Instance-level fusion strategies, which exchange sparse object-centric information, seem more resilient to the perspective shifts from varying altitudes than their dense, BEV-level counterparts. Meanwhile, the findings suggest that resilience to localization errors is directly linked to adaptive data filtering. Methods that selectively fuse information—either through instance-level filtering or scene-level spatial confidence maps—proved more robust, highlighting that successful cooperation requires not just sharing data, but discerning which data to trust and fuse.

These findings indicate clear directions for future research. Efforts should focus on developing altitude-adaptive and scale-aware fusion mechanisms capable of handling dynamic aerial viewpoints. Advancing sparse fusion methods to strike a better balance between performance and communication costs and creating dynamic trust mechanisms to weigh or filter erroneous signals will be critical. Addressing these challenges is essential for developing robust aerial-ground cooperative systems that can be reliably deployed in unpredictable, real-world conditions.

Furthermore, our benchmark could be extended in several key areas. We encourage the exploration of more advanced late-fusion strategies (Chiu et al. 2024), which may boost performance at a minimal communication cost. We also see value in benchmarking perception models specifically designed for the aerial domain, which may provide more robust features for fusion. Finally, future robustness analyses should investigate the impact of diverse weather conditions, evaluate performance under fair bandwidth-constrained scenarios, and incorporate asynchrony-robust methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China for the Science Fund for Creative Research Groups (No. 52221005) and the Key Project (No. 52131201).

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11618–11628.

Chiu, H.-K.; Wang, C.-Y.; Chen, M.-H.; and Smith, S. F. 2024. Probabilistic 3D Multi-Object Cooperative Tracking for Autonomous Driving via Differentiable Multi-Sensor Kalman Filter. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 18458–18464.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, 1–16. PMLR.

Gao, X.; Wu, Y.; Luo, X.; Wu, K.; Chen, X.; Wang, Y.; Liu, C.; Zhou, Y.; and Tu, Z. 2025. AirV2X: Unified Air-Ground Vehicle-to-Everything Collaboration. arXiv:2506.19283.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision Meets Robotics: The KITTI Dataset. *International Journal of Robotics Research*, 32(11): 1231–1237.

Hao, R.; Fan, S.; Dai, Y.; Zhang, Z.; Li, C.; Wang, Y.; Yu, H.; Yang, W.; Yuan, J.; and Nie, Z. 2024. RCooper: A Real-World Large-Scale Dataset for Roadside Cooperative Perception. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22347–22357.

Hou, Y.; Zou, B.; Zhang, M.; Chen, R.; Yang, S.; Zhang, Y.; Zhuo, J.; Chen, S.; Chen, J.; and Ma, H. 2025. AGC-Drive: A Large-Scale Dataset for Real-World Aerial-Ground Collaboration in Driving Scenarios. arXiv:2506.16371.

Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; and Chen, S. 2022. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems*, volume 35, 4874–4886.

Kuhn, H. W. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97.

Kutila, M.; Kauvo, K.; Pyykönen, P.; Zhang, X.; Martinez, V. G.; Zheng, Y.; and Xu, S. 2021. A C-V2X/5G Field Study for Supporting Automated Driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, 315–320.

Li, Y.; Ma, D.; An, Z.; Wang, Z.; Zhong, Y.; Chen, S.; and Feng, C. 2022a. V2X-sim: Multi-Agent Collaborative Perception Dataset and Benchmark for Autonomous Driving. *IEEE Robotics and Automation Letters*, 7(4): 10914–10921.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022b. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, 1–18. Cham: Springer Nature Switzerland. ISBN 978-3-031-20077-9.

Shah, S.; Dey, D.; Lovett, C.; and Kapoor, A. 2018. AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In *Field and Service Robotics*, 621–635. Cham: Springer International Publishing. ISBN 978-3-319-67361-5.

Song, Z.; Xie, T.; Zhang, H.; Liu, J.; Wen, F.; and Li, J. 2024. A Spatial Calibration Method for Robust Cooperative Perception. *IEEE Robotics and Automation Letters*, 9(5): 4011–4018.

Tian, P.; Cheng, P.; Wang, Y.; Wang, Z.; Wang, Z.; Yan, M.; Yang, X.; and Sun, X. 2024. UCDNet: Multi-UAV Collaborative 3D Object Detection Network by Reliable Feature Mapping. *IEEE Transactions on Geoscience and Remote Sensing*.

Wang, Y.; Cheng, P.; Tian, P.; Yuan, Z.; Zhao, L.; Tian, J.; Wang, W.; Wang, Z.; and Sun, X. 2024a. UVCPNet: A UAV-vehicle Collaborative Perception Network for 3D Object Detection. arXiv:2406.04647.

Wang, Y.; Wang, Z.; Cheng, P.; Tian, P.; Yuan, Z.; Tian, J.; Wang, W.; and Zhao, L. 2025. AVCPNet: An AAV-Vehicle Collaborative Perception Network for 3-D Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–16.

Wang, Z.; Cheng, P.; Chen, M.; Tian, P.; Wang, Z.; Li, X.; Yang, X.; and Sun, X. 2024b. Drones Help Drones: A Collaborative Framework for Multi-Drone Object Trajectory Prediction and Beyond. *Advances in Neural Information Processing Systems*, 37: 64604–64628.

Xu, R.; Xia, X.; Li, J.; Li, H.; Zhang, S.; Tu, Z.; Meng, Z.; Xiang, H.; Dong, X.; Song, R.; Yu, H.; Zhou, B.; and Ma, J. 2023. V2V4Real: A Real-World Large-Scale Dataset for Vehicle-to-Vehicle Cooperative Perception. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13712–13722.

Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; and Ma, J. 2022a. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, 107–124. Cham: Springer Nature Switzerland. ISBN 978-3-031-19842-7.

Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; and Ma, J. 2022b. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In *2022 International Conference on Robotics and Automation (ICRA)*, 2583–2589.

Ye, H.; Sunderraman, R.; and Ji, J. S. 2024. UAV3D: A Large-scale 3D Perception Benchmark for Unmanned Aerial Vehicles. *Advances in Neural Information Processing Systems*, 37: 55425–55442.

Yu, H.; Luo, Y.; Shu, M.; Huo, Y.; Yang, Z.; Shi, Y.; Guo, Z.; Li, H.; Hu, X.; Yuan, J.; and Nie, Z. 2022. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21329–21338.

Yu, H.; Yang, W.; Ruan, H.; Yang, Z.; Tang, Y.; Gao, X.; Hao, X.; Shi, Y.; Pan, Y.; Sun, N.; Song, J.; Yuan, J.; Luo,

P.; and Nie, Z. 2023. V2X-seq: A Large-Scale Sequential Dataset for Vehicle-Infrastructure Cooperative Perception and Forecasting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5486–5495.

Yu, H.; Yang, W.; Zhong, J.; Yang, Z.; Fan, S.; Luo, P.; and Nie, Z. 2025. End-to-End Autonomous Driving Through V2X Cooperation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9): 9598–9606.

Zhong, J.; Wang, J.; Xu, J.; Li, X.; Nie, Z.; and Yu, H. 2025. CoopTrack: Exploring End-to-End Learning for Efficient Cooperative Sequential Perception. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 26954–26965.

Zhou, Y.; Quang, L.; Nieto-Granda, C.; and Loianno, G. 2024. CoPeD-advancing Multi-Robot Collaborative Perception: A Comprehensive Dataset in Real-World Environments. *IEEE Robotics and Automation Letters*, 9(7): 6416–6423.