Supplementary Materials (Appendix) for INSTINCT: Instance-Level Interaction Architecture for Query-Based Collaborative Perception

Yunjiang Xu[1], Lingzhi Li[1]*, Jin Wang[2]*, Yupeng Ouyang[1], Benyuan Yang[2]
[1]School of Computer Science and Technology, Soochow University
[2]School of Future Science and Engineering, Soochow University
{yjxu95, ouyangchina1}@gmail.com, {lilingzhi, wjin1985, byyang}@suda.edu.cn

1. More details

1.1. Implement details

INSTINCT employs SECOND [6] as the 3D backbone and uses ConQueR [8] as the single-agent baseline to provide instance-level features. We set the number of layers in both the encoder and decoder within ConQueR to 3. Moreover, in the final layer of the decoder, the Focal Loss used for classification supervision is replaced with MAL Loss [4], with the corresponding γ set to 1. In the subsequent filtering operation, we set the threshold to 0.1, which is identical to the λ value used in our Dual-branch Detection Routing (DDR).

After the collaborative output head, all instances are aggregated, and a unified detection head, which is composed of a 3-layer FFN, is used to generate the final output. Notably, for classification supervision at this stage, we also employ MAL Loss configured the same way as in the decoder's final layer to penalize detection outputs that exhibit low IoU yet high confidence scores.

Both single-agent and collaborative outputs are supervised. As described in DETR [1], we utilize Hungarian matching for loss computation, and the number of queries is initialized to 300.

Technical clarification of Co-GT Sampling strategy: Co-GT Sampling constructs an object database $D = \{(P^{(i)}, L^{(i)})\}$, where $P^{(i)}$ and $L^{(i)}$ denote the i-th cropped object point cloud and its label. A subset $\mathcal{O} = \{o^{(j)}\} \subset D$ is selected such that $\mathrm{IoU}(o^{(j)}, o^{(k)}) = \mathrm{IoU}(o^{(j)}, \mathrm{GT}) = 0$ for all $j \neq k$, ensuring no overlap with other samples or existing GTs. For each agent a, its point cloud P_a and labels L_a are transformed to the ego frame via $T_{e\leftarrow a}$, merged with $\mathcal{O}$ and corresponding labels, then inverse-transformed by $T_{a\leftarrow e}$ to produce the augmented pair $(P_a^{\mathrm{aug}}, L_a^{\mathrm{aug}})$: $(P_a, L_a) \xrightarrow{T_{e\leftarrow a}} (P_a \cup \mathcal{O},\, L_a \cup L_{\mathcal{O}}) \xrightarrow{T_{a\leftarrow e}} (P_a^{\mathrm{aug}}, L_a^{\mathrm{aug}})$. Legal samples $\mathcal{O}$ are thus aligned in the ego frame, shared across agents, and restored to local frames for conflict-free, consistent augmentation.

1.2. Experiments details

Sufficient repetition of experiments (mean ± std): All results in the manuscript are averaged over 5 runs with random seeds. We further reran 10 times on DAIR-V2X, yielding AP@0.5 = 0.8287 $\pm$ 0.0048 and AP@0.7 = 0.7539 $\pm$ 0.0067, confirming INSTINCT's consistent superiority over all baselines.

Runtime and Bandwidth Details: Excluding the 3D backbone, collaborative modules incur inference time of 110/120/62 ms (INSTINCT/V2X-ViT/Where2comm) on an A100, with total time of 175/110/77 ms. This indicates that INSTINCT's backbone is the primary speed bottleneck, and its collaborative modules remain inefficient. Bandwidth averages 16 KB/frame across datasets; on DAIR-V2X it is 12 KB (median 11, range 0–36, variance 8), scaling linearly with object count, while other methods grow quadratically with sensing range.

Comparison to TransIFF [2] & QUEST [3] & ACCO [7]: Although TransIFF achieves low bandwidth, it falls short of mainstream SOTA performance and is limited to vehicle-infrastructure collaboration. Our reproduction using the same single detector as INSTINCT yields AP@0.5/0.7 of 0.6919/0.5879 on DAIR-V2X (higher than reported), with a mean bandwidth of 13.58, same as INSTINCT due to identical detector use. Thus, aside from bandwidth, TransIFF underperforms compared to SOTA methods. QUEST and ACCO also adopt instance-level interaction, but are designed for camera-based perception, whereas INSTINCT focuses on LiDAR. Although INSTINCT outperforms them, direct comparison is unfair due to the modality gap.

Performance gain: collaboration or single detector? We addressed this in Appendix Sec.2, comparing against two baselines: late fusion with the best single detector, and INSTINCT using the same weights. INSTINCT outperforms late fusion by 12.17% at AP@0.5 and 25.95% at AP@0.7, demonstrating its effectiveness in calibrating de-

*Corresponding author.

tections. We replace the single detector with Voxel-DETR (whose mAP@0.7 is ∼5% lower than ConQueR) to test INSTINCT on DAIR-V2X. The AP@0.7 drops slightly to 0.7303 (by 2%), while the communication cost increases to 13.69. It indicates that INSTINCT is robust to weaker detectors, though it needs higher communication overhead.

2. Comparison with other 3D Detectors

Advanced single-vehicle models provide sufficiently accurate instance features for INSTINCT. To demonstrate INSTINCT's ability to effectively fuse these instance features, we compared its performance with other state-of-the-art 3D detectors that employ direct late fusion. Additionally, we evaluated the impact of using MAL loss versus quality prediction on detection performance. In the quality prediction approach, an extra 3-layer FFN is used to predict the IoU, and the IoU between positively matched samples and the ground truth (obtained through Hungarian matching) serves as the supervision signal.

Based on the results in Tab. 1, the performance of the INSTINCT model across different datasets exhibits the following characteristics:

1. **Performance Comparison on the DAIR-V2X Dataset.** INSTINCT's late fusion performance is significantly superior to that of ConQueR and SEED . This demonstrates that under the complex data distributions encountered in real-world scenarios (e.g., variations in illumination, occlusions), the instance fusing mechanism in INSTINCT can more effectively integrate information from multiple agents. In contrast, traditional late fusion methods lack effective calibration capabilities, struggle to handle the noise interference present in real environments.
2. **Analysis of the V2XSet Simulation Dataset.** Although INSTINCT achieves the best performance in the intermediate fusion model, its performance is slightly lower than the late fusion results of ConQueR and SEED [5] in V2XSet. We attribute this to two factors:
   - **Simplicity of the Data Distribution.** As a simulated dataset, the distribution of 3D detection targets in V2XSet is highly idealized, resulting in near-saturation of prediction accuracy among the agents. In such cases, maintaining the original predictions can actually be the optimal strategy.
   - **Model Training Challenges.** INSTINCT is designed to learn "how to fuse" instance features through training, but the strong consistency of "unchanged" targets in simulated data makes it difficult for the model to converge to the desired state.
   - **Practical Implication.** Since collaborative perception is ultimately applied in real autonomous driving scenarios (e.g., the complex road tests in DAIR-V2X), the advantages of INSTINCT on real-world data hold greater engineering value.
3. **Trade-offs Between the Loss Function and the Quality Prediction Module.**
   - **MAL Loss.** This loss function consistently boosts performance in both the multi-agent baseline models and INSTINCT, validating its effectiveness in enhancing instance selection through adversarial learning.
   - **Quality Prediction (IoU).** On the V2XSet dataset, the simplicity of the targets results in an 1.5% gain. However, on the DAIR-V2X dataset, the limited data scale leads to prediction biases that cause a performance drop. In the current approach, we rely solely on MAL Loss for quality-aware filtering. Future work will focus on improving the robustness of IoU prediction through a pre-training and fine-tuning strategy or uncertainty modeling.

This analysis highlights the strengths of INSTINCT, particularly in real-world scenarios, and outlines potential avenues for further enhancing the model's robustness and performance.

3. Effectiveness of CDA and GDA

To more intuitively illustrate the roles of the two modules, CDA and GDA, in Cross-Agent Local Instance Fusion (CALIF), we provide a visualization of the fused instances from the ego and agents. As shown in Fig. 1(a), after applying CDA, the numerical distributions of the ego and agent instances become more similar. Following GDA, both instances undergo self-calibration based on each other's information. Notably, ego instance features clearly acquire additional information from the agent instance features. In Fig. 1(b), we visualize the weights corresponding to GDA, which indicate the degree of mutual attention between the mixed instances. The weight matrix generally appears symmetric, suggesting a tight correspondence between the instance features. It is important to note that similar colors do not imply identical values, but rather comparable attention levels. The attention weights provided by GDA facilitate effective local fusion of the fused instance features.

4. More Visualization Results

We provide additional visual comparisons of INSTINCT on V2XSet and V2V4Real in the Fig. 2 and Fig. 3. In addition to BEV detection results, we also present visualizations of the 3D detection result. All visualizations indicate that INSTINCT exhibits stronger calibration capability and more stable performance compared to state-of-the-art intermediate fusion methods.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

Model	Fusion Method	Quality Aware Method	DAIR-V2X	V2XSet
			AP@0.5/0.7	
ConQueR	Late Fusion	\	0.7252/0.5840	0.9310/0.8974
		MAL	0.7302/0.5978	0.9489/0.9085
		MAL + Quality Prediction	0.7213/0.5717	**0.9593/0.9242**
SEED	Late Fusion	\	0.7332/0.5961	0.9377/0.9064
INSTINCT	Instance	\	0.8077/0.7415	0.9123/0.8688
		MAL	**0.8191/0.7529**	0.9229/0.8731
		MAL + Quality Prediction	0.8111/0.7401	0.9272/0.8822

Table 1. **Comparison with other 3D Detectors.**

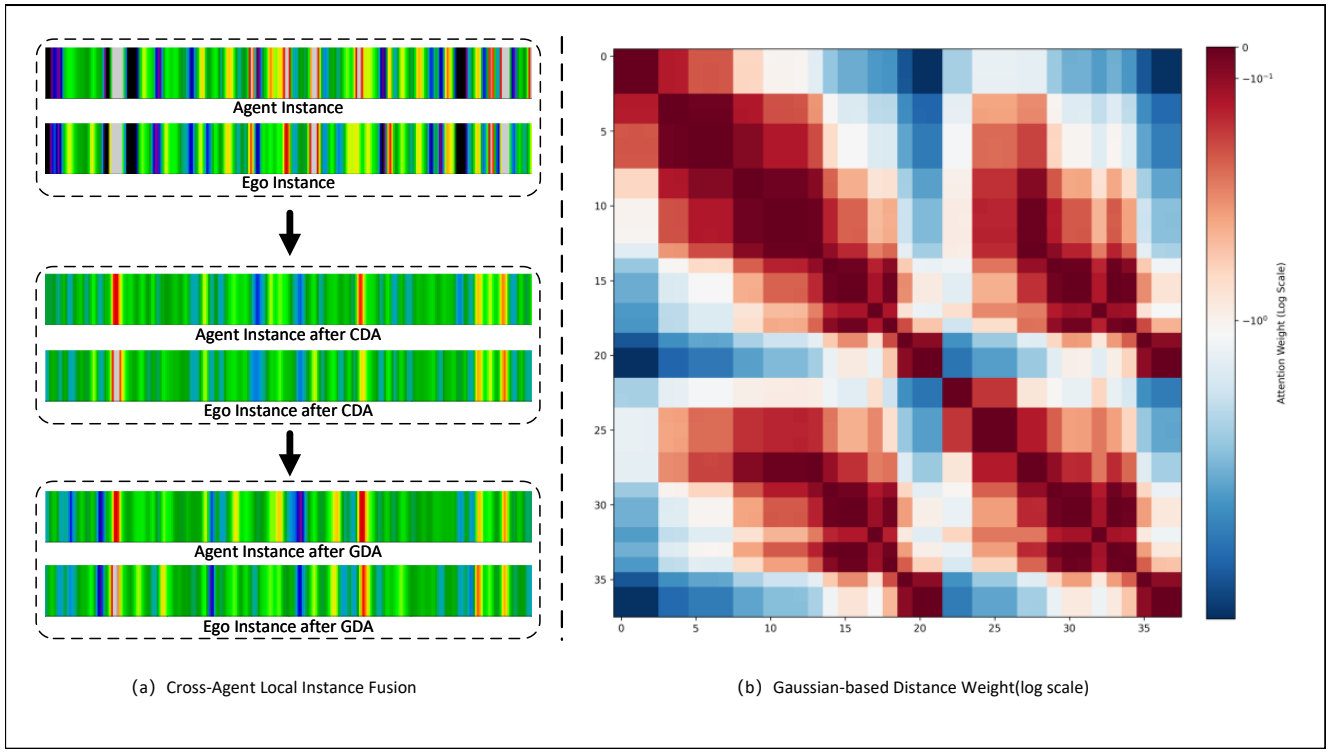

Figure 1. **Visualization of effectiveness of CDA and GDA.** (a) shows the the query after CDA and GDA. (b) shows the relationship between Gaussian distance and weight.

[2] Ziming Chen, Yifeng Shi, and Jinrang Jia. Transiff: An instance-level feature fusion framework for vehicle-infrastructure cooperative 3d detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18205–18214, 2023. 1

[3] Siqi Fan, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Quest: Query stream for practical cooperative perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18436–18442. IEEE, 2024. 1

[4] Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. Deim: Detr with improved matching for fast convergence. *arXiv preprint arXiv:2412.04234*, 2024. 1

[5] Zhe Liu, Jinghua Hou, Xiaoqing Ye, Tong Wang, Jingdong Wang, and Xiang Bai. Seed: A simple and effective 3d detr in point clouds. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 2

[6] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1

[7] Kang Yang, Tianci Bu, Lantao Li, Chunxu Li, Yongcai Wang, and Deying Li. Is discretization fusion all you need for collaborative perception? In *arXiv preprint arXiv:2503.13946*, 2025. 1

[8] Benjin Zhu, Zhe Wang, Shaoshuai Shi, Hang Xu, Lanqing Hong, and Hongsheng Li. Conquer: Query contrast voxel-detr for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2023. 1

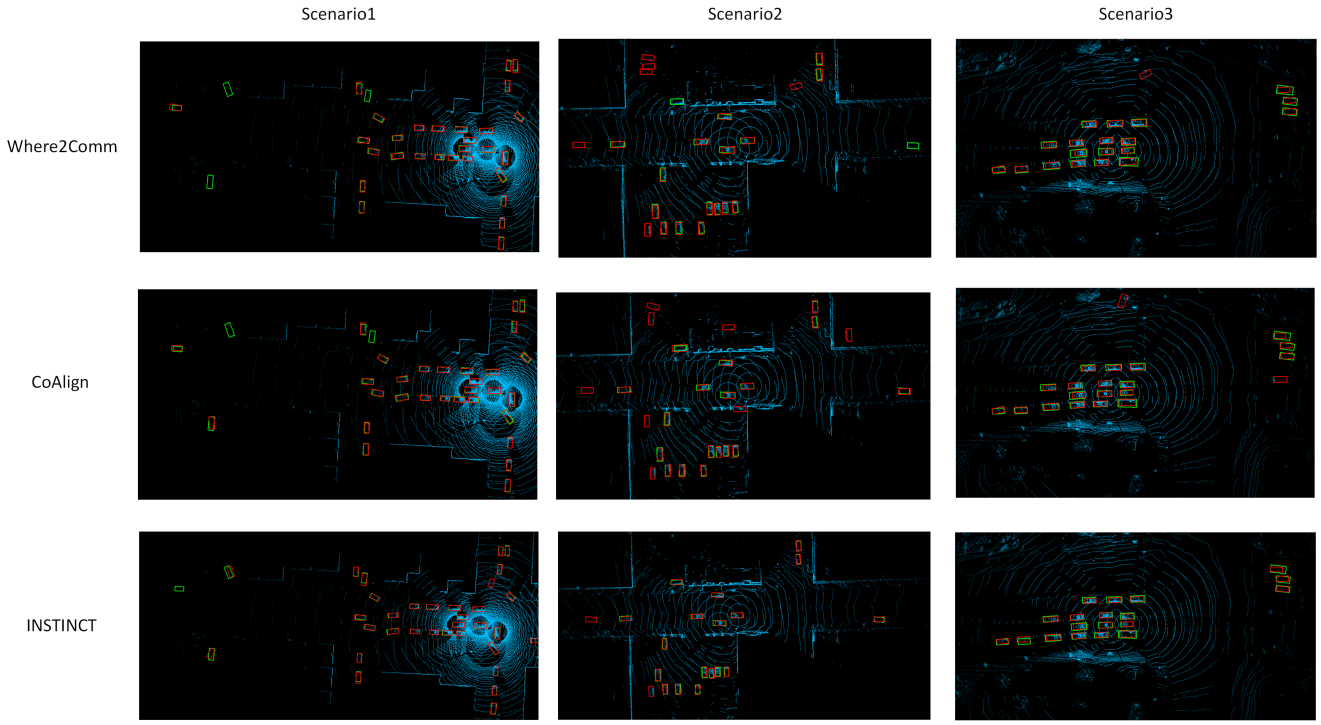

Figure 2. **Comparison of BEV Visualization of Different Models in Different Scenarios.**

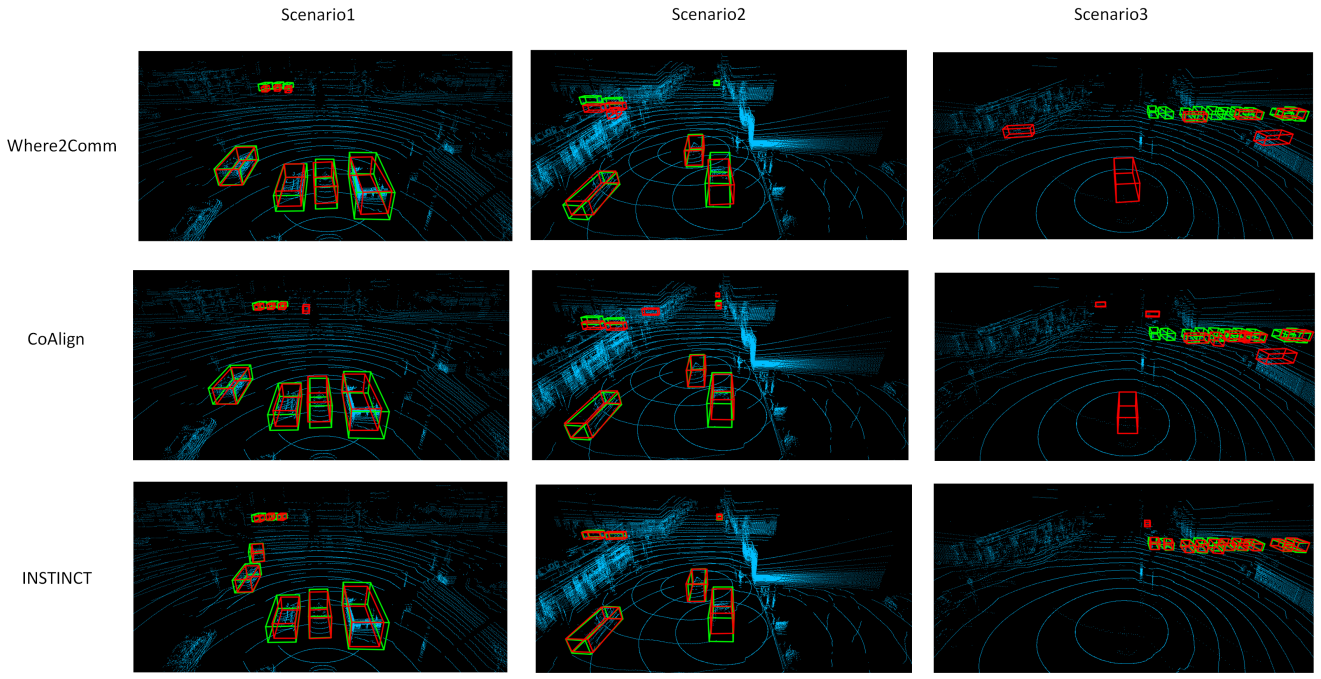

Figure 3. **Comparison of 3D Visualization of Different Models in Different Scenarios.**