WhisperNet: A Scalable Solution for Bandwidth-Efficient Collaboration

Supplementary Material

1. Communication Cost Quantification

Fair comparisons in collaborative perception necessitate a comprehensive quantification of communication cost. Total inter-agent data transmission typically comprises two main components: the **primary data payload** and **protocol-specific overhead**.

The primary payload comprises the intermediate features extracted by each agent's local perception network, constituting the bulk of the transmitted data. The protocol overhead encompasses all non-payload data, which varies significantly by method. This overhead includes essential auxiliary information (e.g., 6-DoF pose) as a fixed cost, alongside variable coordination costs. For instance, a one-shot broadcast has minimal coordination cost, while protocols involving multi-round Query/Key exchanges can be costly. Our proposed method, WhisperNet, utilizes a highly efficient request-response scheme, which incurs only a micro-scale coordination cost for this interaction.

For WhisperNet, this communication process is formalized as follows. The total cost $Size(\mathbb{M}_i)$ for agent i is defined by its constituent parts:

$$Size(\mathbb{M}_i) = Size(I_i) + Size(\Pi_i) + Size(M_i) \quad (1)$$

where these components map directly to our two-part cost structure:

- **Primary payload ($Size(M_i)$):** The size of the final sparse feature message transmitted by agent i, generated according to the received allocation plan Π_i.
- **Protocol overhead ($Size(I_i) + Size(\Pi_i)$):** The sum of our two micro-scale coordination components: (1) $Size(I_i)$, the size of the compact importance map generated and broadcast by agent i to the coordinator; and (2) $Size(\Pi_i)$, the size of the communication allocation plan that agent i receives back from the coordinator.

In all our experiments, the reported communication rates and bandwidth usage are calculated based on this comprehensive definition of $Size(\mathbb{M}_i)$, ensuring a fair and rigorous comparison against baseline methods.

As illustrated in Fig. 1, the communication cost breakdown on OPV2V reveals the efficiency of our protocol. The coordination overhead is fixed and minimal, comprising only a 94.5 KB transmitted Importance Map and a 47.25 KB received Allocation Plan. The Feature Message payload, in contrast, scales dynamically with the communication rate. This dynamic scaling shifts the component ratio between the Feature Message, Importance Map, and Allocation Plan from approximately 256:2:1 at a 50% rate to just 5.12:2:1 at a 1% rate. Crucially, this analysis demonstrates the protocol's high efficiency. Even at the most constrained 1% rate, which amounts to only 0.24 MB total, the feature payload remains the dominant component at 1.7× the size of the combined coordination overhead. This ultimately confirms that our protocol's minimal, fixed metadata overhead acts as a highly effective lever, enabling substantial reductions in the primary data payload by transmitting only the most critical scene features.

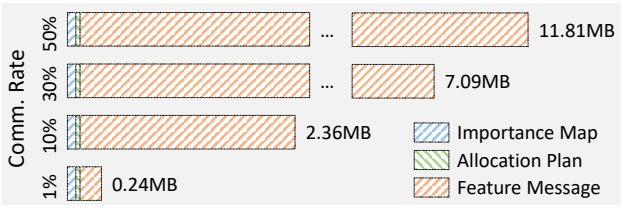

Figure 1. Quantitative breakdown of communication cost components on OPV2V.

2. Extended Discussion on Communication Paradigms

As discussed in the main paper's Related Work section, our method introduces a global coordination paradigm. This section provides further visual and textual elaboration, using Fig. 2 to visually contextualize our contribution.

(1) Compression and Reconstruction. As visualized in Fig. 2(a), earlier schemes [2, 4] employ holistic feature compression. These static, fixed-rate encoders lack adaptability to dynamic scene complexity or bandwidth, failing to optimally allocate resources.

(2) Ego-centric or Region-based Transmission. Fig. 2(b) illustrates a more recent paradigm [1, 3, 6] focused on spatial selection. While spatially adaptive, these methods remain sender-centric or rely on pairwise optimization. This design inherently risks redundant transmissions for overlapping regions or creating perceptual gaps, as it lacks a system-wide, global perspective.

(3) Global Coordination via Joint Metadata Interaction. In sharp contrast, our WhisperNet framework, shown in Fig. 2(c), implements true global coordination. By leveraging joint metadata interaction, the receiver acts as a central coordinator, enabling all agents to *jointly* determine a complementary and non-redundant feature contribution from *all* participants for *every* region. This multi-vehicle decision process maximizes bandwidth efficiency and preserves scene consistency.

Distinction from How2comm. While How2comm [5] also

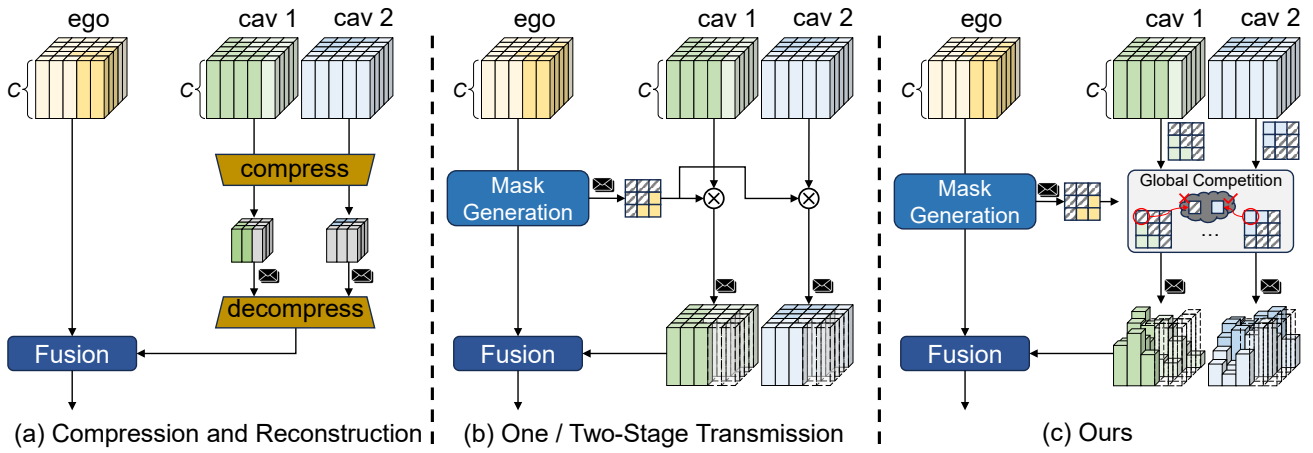

Figure 2. Quantitative breakdown of communication cost components on OPV2V.

addresses channel redundancy, WhisperNet differs in several fundamental aspects: (i) How2comm's filter still operates within a bilateral, request-response paradigm. As analyzed in (2), this pairwise optimization lacks system-wide awareness, risking redundant transmissions or perceptual gaps. In contrast, WhisperNet implements a truly global coordination scheme, achieving low-bandwidth, high-fidelity perception by allocating resources across all agents, patches, and channels. (ii) How2comm employs a CBAM-like module to perform holistic channel feature selection for its own local data. This local gating mechanism, however, operates in isolation; it can neither align heterogeneous channels from different agents nor resolve data conflicts in bilateral exchanges. Our module, in contrast, explicitly performs cross-agent channel-level alignment before feature fusion. (iii) How2comm attempts to compensate for these drawbacks with a complex fusion module, leading to a significant surge in parameters (see Tab. 1 in the main text). In contrast, WhisperNet's more elegant design is not only more computationally efficient but also functions as a plug-and-play module that can seamlessly reduce a model's bandwidth while maintaining performance.

This globally coordinated design highlights the novelty of WhisperNet in achieving system-wide collaborative perception.

3. More Experiments

3.1. Backbone Generalizability Analysis

To validate the generalizability of our proposed WhisperNet framework, we conducted an additional supplementary experiment. All experiments in the main paper are based on PointPillars as the feature extraction backbone. As shown in Tab. 1, we uniformly replaced the backbone of all methods with VoxelNet and evaluated the performance on OPV2V.

WhisperNet continues to demonstrate superior performance and bandwidth efficiency when paired with VoxelNet. As shown in the table, our method achieves a peak performance of 0.8701 AP@0.7 using only 60% bandwidth, surpassing all baselines regardless of their bandwidth allowance. More strikingly, even when bandwidth is compressed to just 20%, WhisperNet's resulting performance of 0.8695 AP@0.7 still outperforms all competing methods at any bandwidth, including CORE at 80% with 0.8566 AP@0.7 and AttFuse at 100% with 0.8530 AP@0.7.

Table 1. Generalizability Study on VoxelNet Backbone.

Methods	Backbone	Bandwidth	AP@0.5	AP@0.7
Single	VoxelNet	100%	0.6884	0.5267
Early		100%	0.8838	0.7872
Late		100%	0.8012	0.7386
F-cooper		100%	0.8765	0.7881
AttFuse		100%	0.9094	0.8530
CORE		80%	0.9087	0.8566
Where2comm		80%	0.8638	0.7650
		60%	0.8616	0.7632
		40%	0.8534	0.7513
		20%	0.8295	0.7233
		10%	0.8084	0.6951
WhisperNet		80%	0.9109	0.8697
		60%	0.9117	0.8701
		40%	0.9109	0.8698
		20%	0.9108	0.8695
		10%	0.8864	0.8569
		5%	0.8447	0.8193

This result provides strong evidence that the core advantages of WhisperNet are not dependent on a specific backbone architecture, highlighting the powerful versatility of our method as a plug-and-play module.

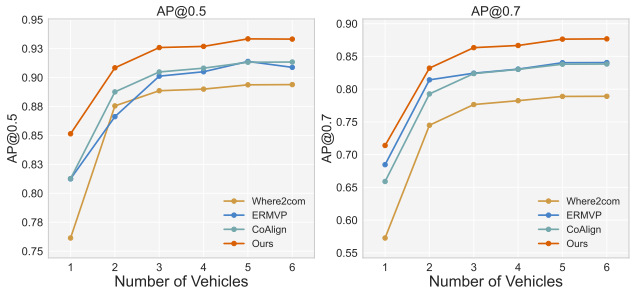

Figure 3. Performance comparison with varying numbers of collaborating agents on OPV2V.

Table 2. Detailed ablation study on the joint impact of the number of collaborating agents (CAVs) and the communication rate on the OPV2V dataset. The table reports Average Precision at IoU thresholds of 0.5 and 0.7.

AP@0.5					
CAVs/Rate	50%	40%	30%	10%	5%
5	0.9337	0.9334	0.9331	0.9323	0.9271
4	0.9276	0.9274	0.9269	0.9265	0.9206
3	0.9267	0.9261	0.9259	0.9247	0.9176
2	0.9092	0.9087	0.9084	0.9063	0.8964
1	0.8514	0.8511	0.8493	0.8411	0.8289
AP@0.7					
CAVs/Rate	50%	40%	30%	10%	5%
5	0.8774	0.8764	0.8764	0.8672	0.8511
4	0.8668	0.8667	0.8660	0.8625	0.8430
3	0.8634	0.8634	0.8624	0.8579	0.8366
2	0.8333	0.8321	0.8316	0.8250	0.8007
1	0.7140	0.7132	0.7121	0.6939	0.6631

3.2. Analysis on the Number of Agents

To investigate the scalability of our proposed framework, we evaluate its performance by varying the number of collaborating agents from 1 to 6. The results, presented in Fig. 3, show that the performance of all collaborative methods generally improves with an increasing number of agents. This trend validates the core benefit of multi-agent perception in mitigating occlusion and expanding perceptual range. Crucially, our proposed method, WhisperNet, consistently outperforms all strong baselines, including CoAlign, Where2com, and ERMVP, across all tested agent counts. Notably, the performance gap between WhisperNet and these other methods remains significant as more agents are introduced. This result demonstrates the superior scalability of our framework, indicating that its joint spatial-channel selection and fusion mechanism more effectively leverages the information from additional agents to achieve higher perception accuracy.

3.3. Impact of Agent Scale and Bandwidth

Tab. 2 presents a ablation study that evaluates the performance of WhisperNet under a matrix of conditions, varying both the number of collaborating agents and the communication data rate. The results demonstrate two key trends. First, our method scales with the number of agents, as performance consistently improves across all communication rates when more vehicles collaborate. Second, WhisperNet exhibits remarkable robustness to bandwidth constraints. For any given number of agents, the performance degradation is very graceful as the communication rate is drastically reduced. For instance, with 5 CAVs, the AP@0.7 score only drops from 0.8774 to 0.8511 even when the data rate is cut by 90% (from 50% to 5%). This comprehensive analysis validates the scalability and high communication efficiency of our proposed approach.

4. Heatmap Visualization of Features

Fig. 4 provides a qualitative comparison of the feature maps transmitted by different communication-efficient methods, offering insight into what information each strategy prioritizes and preserves under bandwidth constraints. Methods based on holistic feature compression, such as CORE, tend to produce diffuse feature maps where high-frequency details are lost across the entire scene. The resulting representation lacks the sharpness required to distinguish fine-grained object boundaries and environmental cues. In contrast, purely spatial selection methods like Where2comm and ERMVP focus on transmitting high-fidelity features from sparse, disjoint regions deemed important. However, this aggressive spatial pruning results in a significant loss of contextual information. As visualized, the background, road layout, and spatial relationships between objects are largely discarded, which can be detrimental for downstream tasks that require a comprehensive understanding of the scene beyond just foreground objects. WhisperNet demonstrates a superior approach by jointly optimizing spatial and channel information. Even at an extreme 10x compression rate, it preserves a coherent, holistic view of the scene. The max projection remains stable compared to the 2x version, indicating that the most salient channel features are consistently transmitted regardless of bandwidth pressure. While the mean projection shows a graceful degradation, it still retains far more environmental context than the baselines. This ability to maintain both key object details and overall scene structure substantiates our method's effectiveness for robust, multi-task collaborative perception.

References

[1] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative

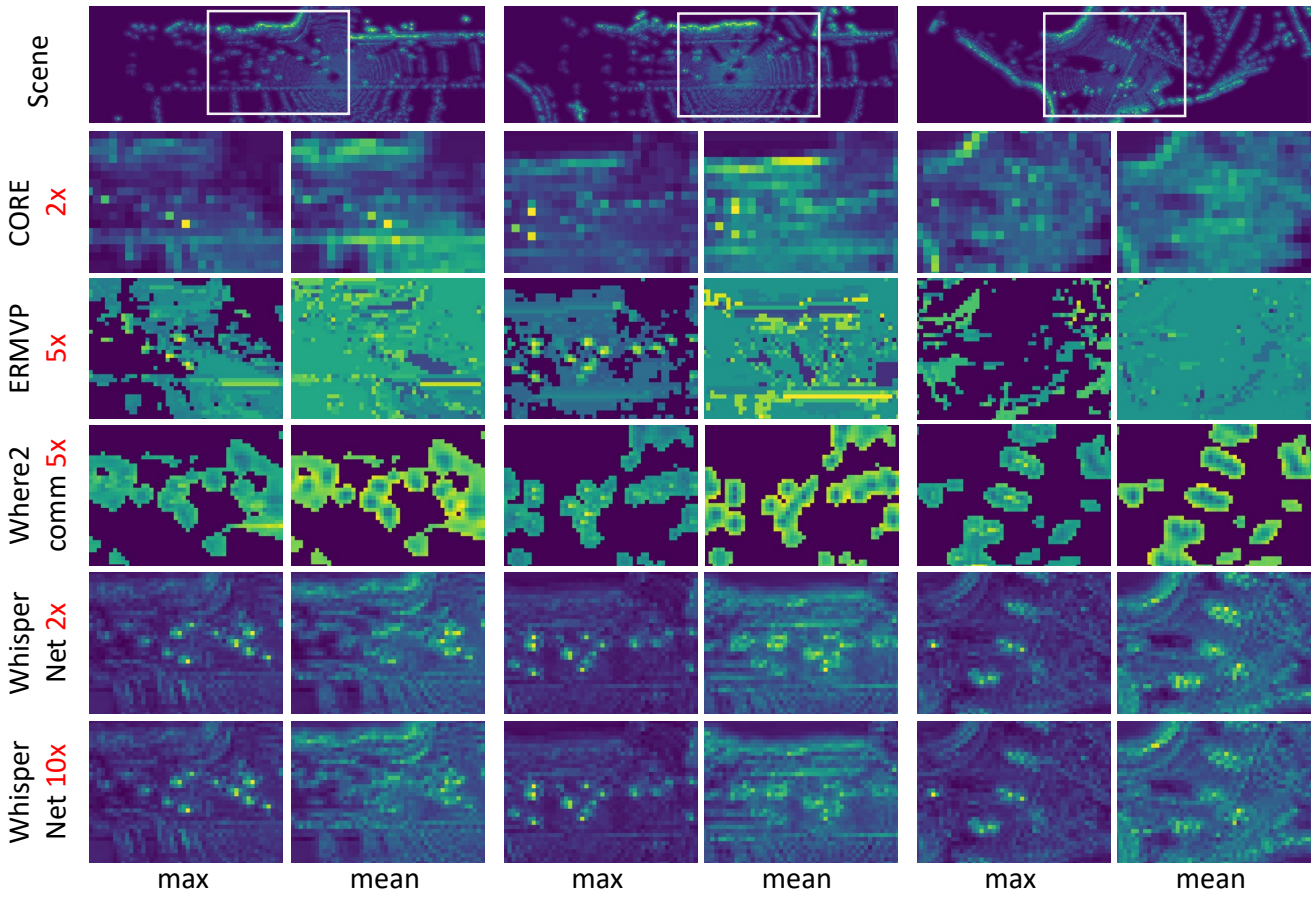

Figure 4. Qualitative Comparison of Transmitted Feature Maps. Visualization of transmitted feature maps under different communication-efficient strategies. We compare our method, WhisperNet, at high (2x) and extreme (10x) compression rates against several baselines: CORE, ERMVP, and Where2comm. We show both max and mean projections across channels to illustrate the richness of the transmitted information.

perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022. 1

[2] Binglu Wang, Lei Zhang, Zhaozhong Wang, Yongqiang Zhao, and Tianfei Zhou. Core: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8710–8720, 2023. 1

[3] Junhao Xu, Yanan Zhang, Zhi Cai, and Di Huang. Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 6834–6843, 2025. 1

[4] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 1

[5] Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36:25151–25164, 2023. 1

[6] Jingyu Zhang, Kun Yang, Yilei Wang, Hanqi Wang, Peng Sun, and Liang Song. Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12575–12584, 2024. 1