Table 1. **Space-time complexity and inference time comparison.** Limited by the rebuttal time, we select recent representative SOTA models for comparison. We adopt two mainstream metrics to evaluate the space-time complexity, including the Parameter Number and FLOPs. Fps is used to compare the inference time. Intuitively, How2comm outperforms the previous SOTAs on all five datasets and remains competitive regarding space-time complexity and inference time. The results prove that How2comm can achieve a better trade-off among training cost, inference time, and perception performance than most SOTA methods.

Metrics	V2X-ViT	CoBEVT	Where2comm	How2comm (Ours)
Parameter Number (M) ↓	18.23	16.76	12.08	16.15
FLOPs (G) ↓	213	152	106	148
Fps (DAIR-V2X) ↑	9.43	19.36	14.91	16.07
Fps (V2XSet) ↑	6.87	14.11	11.82	12.61
Fps (OPV2V) ↑	6.86	13.21	12.69	12.34
Fps (OPV2V Culver City) ↑	8.74	17.67	14.66	15.93
Fps (V2V4Real) ↑	13.37	25.28	20.58	22.45
AP@0.5/0.7 (DAIR-V2X) ↑	51.68/39.97	56.08/41.45	59.34/43.53	**61.95/46.77**
AP@0.5/0.7 (V2XSet) ↑	76.95/58.14	76.46/57.66	76.80/59.84	**78.66/61.57**
AP@0.5/0.7 (OPV2V) ↑	80.61/66.42	81.59/67.50	82.75/67.29	**84.39/69.37**
AP@0.5/0.7 (OPV2V Culver City) ↑	73.65/55.83	74.16/56.79	72.90/56.31	**75.83/58.07**
AP@0.5/0.7 (V2V4Real) ↑	58.37/30.41	60.17/31.25	59.53/30.64	**61.78/33.26**

Table 2. **Ablation study results of candidate designs and strategies.** "w/o" means without. We set both $\mathcal{E}$ and $\mathcal{C}$ to 1 and use the same attention module in both branches for the ablation study of decoupled design. For the complete module ablation, we remove the FDC module and replace the MIC module with a compression-based method. Then we replace the STCFormer with a 1×1 convolution-based fusion model. In this fusion model, the same fusion pattern is used to aggregate context information and collaborator-shared features, where features are concatenated on the channel dimension and fused by a 1×1 convolutional network.

Designs/Strategies	DAIR-V2X	V2XSet	OPV2V	Culver City	V2V4Real
	AP@0.5/0.7	AP@0.5/0.7	AP@0.5/0.7	AP@0.5/0.7	AP@0.5/0.7
Full Model	**61.95/46.77**	**78.66/61.57**	**84.39/69.37**	**75.83/58.07**	**61.78/33.26**
Effect of Decoupled Design					
w/o Decoupled Design	59.62/44.75	76.70/59.83	82.85/68.16	73.72/56.81	60.25/32.04
Effect of Complete Module					
w/o MIC	60.37/44.53	76.64/60.03	83.04/67.95	74.26/57.01	59.90/31.85
w/o FDC	58.81/44.16	75.92/59.94	82.92/67.07	73.82/56.54	59.57/31.44
w/o STCFormer	53.76/41.74	73.51/54.72	78.57/64.19	70.53/54.08	56.38/28.26

Table 3. **Detection performance comparison on the V2V4Real dataset.** We employ the experimental setup in the Main Paper, where the localization and heading errors are 0.2m and 0.2°, the delay is 100ms, and the bandwidth limitation is ≈1MB. We implement the compared baselines with the public codebase and report detection performance under the above bandwidth-limited noise settings. Notably, How2comm outperforms existing SOTAs and improves the SOTA performance of AP@0.7 by **6.5%**. How2comm presents significant performance gains on the challenging real-world dataset V2V4Real, further proving its superiority.

Model	V2V4Real Dataset	
	AP@0.5	AP@0.7
No Fusion	39.80	22.00
Late Fusion	50.20	22.40
Early Fusion	52.10	25.80
F-Cooper	55.25	27.75
AttFuse	57.44	29.18
V2VNet	59.06	28.92
V2X-ViT	58.37	30.41
CoBEVT	60.17	31.25
Where2comm	59.53	30.64
How2comm (Ours)	**61.78**	**33.26**