One is Plenty: A Polymorphic Feature Interpreter for Immutable Heterogeneous Collaborative Perception

Supplementary Material

1. Experimental Details

When adapting to a new neighbor agent, trainable parameters include a specific prompt for the neighbor agent and a resizer to align the size of the neighbor features with the ego features. The resizer consists of a max-pooling layer and a 1×1 convolution. Specifically, if the ego feature size is $C_1 \times H_1 \times W_1$ and the neighbor feature size is $C_2 \times H_2 \times W_2$, the number of parameters for the specific prompt is $C_2 \times H_1 \times W_1$, and the number of parameters for the resizer is $C_1 \times C_2$. The trainable parameter numbers for adapting to new neighbor agents with pp8 as the ego agent are shown in Table 1.

2. Additional Ablation Study

We perform ablation experiments on three loss components: style loss (regulating both shared and agent-specific semantics), adversary loss (regulating shared semantics), and single loss (regulating agent-specific semantics), with results detailed in Table 2, following the main paper's settings.

3. Performance of Multi-Modal Fusion

We conduct multi-modal fusion experiments by integrating LiDAR and camera data, using two image encoders, EfficientNet [30] and ResNet [11], and conduct two sets of experiments. In the first set, the base model is trained with the combination of pp8-pp4-EfficientNet, where the ego agent collaborates with new neighbor agents using ResNet in Generalization Phase. In the second set, the base model is trained with pp8-vn4-ResNet, and the ego agent collaborates with new neighbor agents using EfficientNet in Generalization Phase. As shown in Table 3, PolyInter performs well in multi-modal fusion.

Encoder	Feature Size	Parameters (M)		
		Specific Prompt	Resizer	Total
pp8	$256 \times 50 \times 176$			
pp6 [16]	$384 \times 64 \times 256$	3.38	0.10	3.48
pp4	$384 \times 100 \times 352$	3.38	0.10	3.48
vn6 [38]	$128 \times 128 \times 512$	1.13	0.03	1.16
vn4 [38]	$128 \times 200 \times 704$	1.13	0.03	1.16
sd2 [44]	$512 \times 50 \times 176$	4.51	0.13	4.64
sd1 [44]	$512 \times 100 \times 352$	4.51	0.13	4.64

Table 1. Trainable parameter numbers of different encoders.

	pp8-pp4	pp8-sd1	pp8-vn6
PolyInter	**77.2 / 65.8**	**79.4 / 66.3**	**72.8 / 55.1**
-w/o style	75.2 / 62.2	77.2 / 63.5	69.6 / 50.1
-w/o adversary	77.0 / 63.4	79.1 / 63.9	71.8 / 54.0
-w/o single	76.4 / 59.7	78.4 / 61.5	72.2 / 52.0

Table 2. Ablation study of loss functions.

4. New Datasets

4.1. Performance Comparison on V2XSet

An open dataset, V2XSet [34], is used in the comparative experiments. Compared to the OPV2V dataset, V2XSet incorporates vehicle-to-everything cooperation and realistic noise simulation. The experimental results comparing PolyInter with PnPDA [25] and MPDA [36] are presented in Table 4.

4.2. Performance Comparison on DAIR-V2X

We also conduct experiments on the real-world DAIR-V2X [42] dataset. We use pp8-pp4-vn4 combination in the base model training phase. The experimental results comparing PolyInter with PnPDA [25] and MPDA [36] are presented in Table 5.

5. Additional Experiments

5.1. Three-agent Collaborative Perception

We compare the performance of PolyInter with PnPDA [25] and MPDA [36] in the immutable heterogeneous scenario of three-agent collaborative perception, as shown in Table 6. The three collaborating agents are set as three different agent types, in the format of "ego-neb1-neb2". The selected scenarios include pp8-pp4-vn6, pp8-pp4-sd1, and pp8-vn6-sd1. In three-agent collaboration, the ego agent's interpreter separately interprets the heterogeneous features of the two neighbor agents into the ego agent's semantic space. The interpreted neighbor features, combined with the ego features, are fed into the fusion module and the detection head on the ego agent to produce the collaborative perception results. The remaining settings are consistent with those described in Sec. 4.2.

5.2. Two-Stage Interpretation of PnPDA

PnPDA [25] adopts a two-stage strategy, where in practical applications, the neighbor features are first interpreted

	w/ F-cooper Fusion			w/ CoBEVT Fusion		
	Ours	**PnPDA** [25]	**MPDA** [36]	**Ours**	**PnPDA** [25]	**MPDA** [36]
pp8-ResNet [11]	**72.6 / 56.9**	61.9 / 42.5	70.7 / 56.6	**75.4 / 57.9**	68.4 / 46.6	71.4 / 56.5
pp8-EfficientNet [30]	**72.2 / 57.3**	60.1 / 41.0	67.4 / 54.4	**72.6 / 57.5**	67.8 / 46.8	70.2 / 56.3

Table 3. LiDAR + camera performance in AP@0.5/AP@0.7.

Interpreter / Scenarios	w/ F-cooper [7] Fusion			w/ CoBEVT [37] Fusion		
	PolyInter (Ours)	**PnPDA** [25]	**MPDA** [36]	**PolyInter (Ours)**	**PnPDA** [25]	**MPDA** [36]
pp8-pp4* [16] pp8-pp4+ [16]	**84.1 / 71.3** **86.7 / 72.2**	80.6 / 55.0	77.8 / 55.4	**86.4 / 72.2** **86.5 / 72.1**	84.5 / 63.9	84.2 / 70.5
pp8-sd1* [16, 38] pp8-sd1+ [16, 38]	**85.8 / 70.3** **87.9 / 74.8**	83.5 / 63.6	78.4 / 54.0	**87.7 / 76.8** **87.4 / 74.5**	86.9 / 63.3	81.4 / 66.6
pp8-vn6* [16, 44] pp8-vn6+ [16, 44]	**80.5 / 71.3** **84.0 / 62.9**	75.9 / 51.7	69.5 / 50.7	**83.8 / 65.4** **83.7 / 63.6**	79.7 / 51.1	70.7 / 51.7

Table 4. Comparison with PnPDA and MPDA on V2XSet dataset.

Interpreter / Scenarios	w/ F-cooper [7] Fusion			w/ CoBEVT [37] Fusion		
	PolyInter (Ours)	**PnPDA** [25]	**MPDA** [36]	**PolyInter (Ours)**	**PnPDA** [25]	**MPDA** [36]
pp8-pp4 [16]	**65.2 / 38.5**	59.3 / 33.9	64.8 / 32.1	**66.8 / 39.9**	62.9 / 34.4	65.0 / 35.1
pp8-sd1 [16, 38]	**65.0 / 38.4**	63.6 / 32.8	64.5 / 34.2	**67.5 / 40.0**	63.4 / 36.0	65.5 / 35.3
pp8-vn6 [16, 44]	**63.9 / 38.2**	49.1 / 30.0	62.8 / 33.6	**65.9 / 39.6**	63.9 / 34.1	64.8 / 34.6

Table 5. Comparison with PnPDA and MPDA on DAIR-V2X dataset.

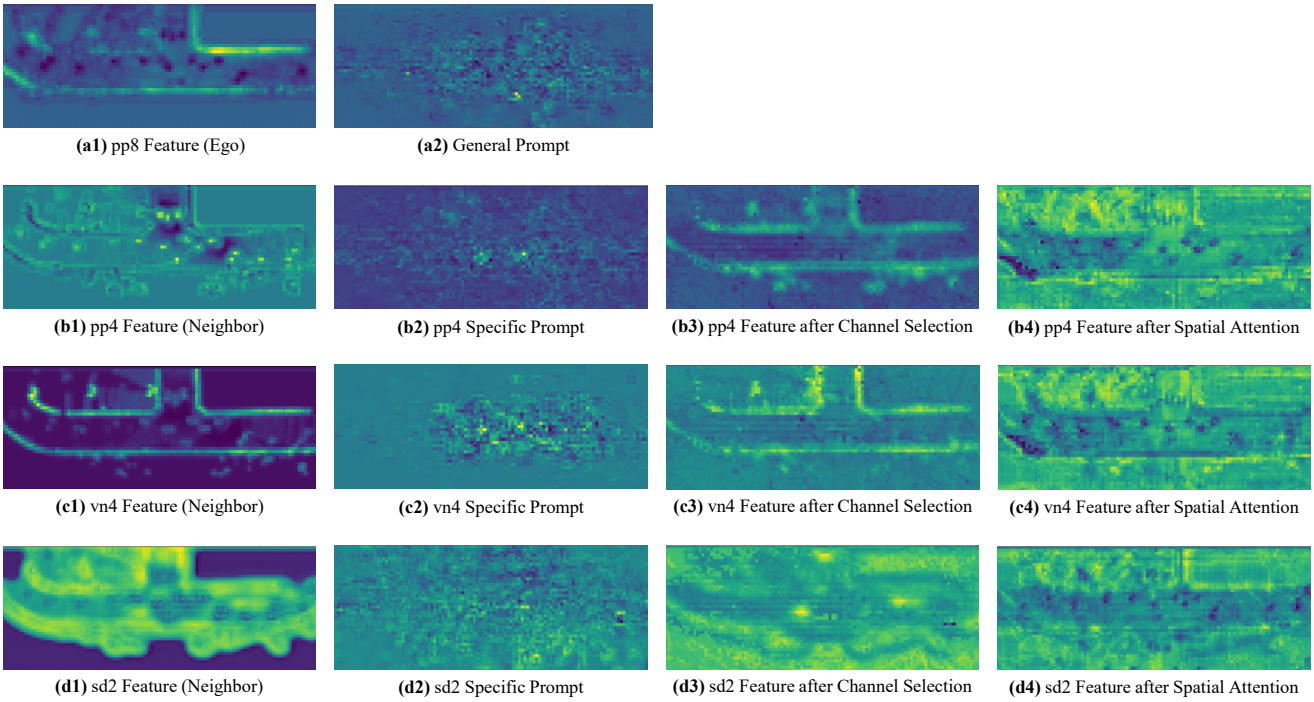
Interpreter / Scenarios	w/ F-cooper [7] Fusion			w/ CoBEVT [37] Fusion		
	PolyInter (Ours)	**PnPDA** [25]	**MPDA** [36]	**PolyInter (Ours)**	**PnPDA** [25]	**MPDA** [36]
pp8-pp4-vn6* [16, 44] pp8-pp4-vn6+ [16, 44]	**77.0 / 61.4** **78.0 / 66.4**	63.5 / 46.4	69.1 / 49.5	**79.2 / 68.6** **78.2 / 67.0**	78.0 / 62.5	72.9 / 57.2
pp8-pp4-sd1* [16, 38] pp8-pp4-sd1+ [16, 38]	**80.3 / 67.2** **79.1 / 69.1**	74.5 / 41.8	73.2 / 45.8	**83.5 / 74.5** **81.1 / 70.9**	79.1 / 65.3	79.6 / 67.0
pp8-vn6-sd1* [16, 38, 44] pp8-vn6-sd1+ [16, 38, 44]	**79.4 / 65.9** **78.8 / 65.9**	68.0 / 48.5	61.4 / 44.9	**80.2 / 70.7** **78.6 / 68.3**	72.0 / 55.2	71.5 / 50.9

Table 6. Comparison with PnPDA and MPDA for three-agent collaborative perception. Our experiments include three heterogeneous scenarios (in the format of "ego-neb1-neb2"): pp8-pp4-vn6, pp8-pp4-sd1, and pp8-vn6-sd1.

into a standard semantic space and then further interpreted into the ego agent's semantic space. The results presented in Sec. 4.2 are from one-stage interpretation, where neighbor features are directly interpreted into the ego agent's semantic space without passing through the standard semantic space. The performance of two-stage interpretation of PnPDA is shown in Table 7. With pp8 as the ego agent and pp4, sd1, and vn6 as the neighbor agents, two agent types, pp4 and vn4, are used as standard semantic spaces, consistent with the settings in [25]. The two-stage interpretation, by passing through the standard semantic space, incurs two stages of semantic loss, which considerably diminishes the collaborative performance.

5.3. Performance of PolyInter in Phase I

The base model is trained with two different encoder combinations in phase I, in the format of "ego-neb1-neb2," including pp8-vn4-sd2 and pp8-pp4-vn4. The performance of the PolyInter base model under these settings is validated,



(a1) pp8 Feature (Ego) **(a2)** General Prompt

(b1) pp4 Feature (Neighbor) **(b2)** pp4 Specific Prompt **(b3)** pp4 Feature after Channel Selection **(b4)** pp4 Feature after Spatial Attention

(c1) vn4 Feature (Neighbor) **(c2)** vn4 Specific Prompt **(c3)** vn4 Feature after Channel Selection **(c4)** vn4 Feature after Spatial Attention

(d1) sd2 Feature (Neighbor) **(d2)** sd2 Specific Prompt **(d3)** sd2 Feature after Channel Selection **(d4)** sd2 Feature after Spatial Attention

Figure 1. Visualization of the ego feature, the general prompt, the specific prompts corresponding to different neighbor agents, and the process of interpreting neighbor features into the ego agent's semantic space.

Standard Semantic	pp4 [16]		vn4 [38]	
Fusion Method	F-cooper [7]	CoBEVT [37]	F-cooper [7]	CoBEVT [37]
pp8-pp4 [16]	77.1 / 51.1	79.2 / 62.3	69.1 / 50.8	73.9 / 57.5
pp8-sd1 [16, 44]	58.7 / 35.0	63.3 / 42.1	63.9 / 41.5	65.4 / 49.7
pp8-vn6 [16, 38]	43.6 / 24.8	57.5 / 30.2	50.8 / 31.1	59.1 / 34.7

Table 7. Two-stage interpretation performance of PnPDA.

Fusion	Combination 1		Combination 2	
F-cooper [7]	pp8-vn4	74.3 / 58.6	pp8-pp4	77.2 / 65.6
	pp8-sd2	81.0 / 66.3	pp8-vn4	74.1 / 61.0
CoBEVT [37]	pp8-vn4	80.2 / 66.0	pp8-pp4	80.2 / 67.0
	pp8-sd2	83.2 / 71.8	pp8-vn4	78.0 / 63.4

Table 8. Performance of PolyInter in phase I. Combination 1 consists of pp8-vn4-sd2, and Combination 2 consists of pp8-pp4-vn4.

with results shown in Table 8.

6. Additional Qualitative Evaluation

As shown in Figure 1, the specific prompts and features for different neighbor agents are visualized. Taking pp8 as the ego agent, pp4, vn4, and sd1 were sequentially selected as neighbor agents. Features of different heterogeneous neighbor agents are matched with distinct specific prompts. The Channel Selection Module reorganizes neighbor features to align with the ego features, while the Spatial Attention Module establishes spatial connections. Finally, all heterogeneous neighbor features are interpreted into the ego agent's semantic space.