One Adapter For All: Towards Generalizable Many-To-One Domain Adaptation In Heterogeneous Collaborative Perception

Bin Lu[1] **Zhewei Fu**[2] **Shaohong Wang**[1] **Zhiyu Xiang**[1] **Hangguan Shan**[1] **Eryun Liu**[1]*

[1]Zhejiang University [2]Zhejiang Dahua Technology

lubin2022@zju.edu.cn eryunliu@zju.edu.cn

Abstract

In autonomous driving, multi-agent collaboration enhances individual perception capabilities through information sharing. However, in real-world applications, differences in models and training data among heterogeneous agents inevitably lead to domain gaps between shared features, bringing challenges to collaboration. To address this, existing methods typically rely on a one-to-one adaptation paradigm, necessitating specific training or fine-tuning for every new agent. This results in a large accumulation of adapters on the vehicle, which is unsustainable for resource-constrained edge devices. To overcome these limitations, we propose UniMTO, a **Uni**fied adaptation framework that enables **M**any-**T**o-**O**ne adaptation across heterogeneous agents using a single, fixed adapter. Specifically, it treats domain adaptation as a pattern transformation task, disentangling domain-invariant content and domain-specific pattern from features, and achieving source-to-target pattern transformation through recombination based on a Mixture-of-Experts (MoE) mechanism. Extensive experiments on V2V4Real and V2X-Real demonstrate that UniMTO achieves superior generalization, and exhibits strong zero-shot adaptation capabilities, outperforming all state-of-the-art methods, enabling many-to-one domain adaptation with a single, fixed adapter.

1 Introduction

Collaborative perception [1] has gained significant attention in recent years. Recent studies have demonstrated that by sharing visual information through vehicle-to-everything (V2X) communication technology, multi-agent perception systems can significantly improve the performance of single-agent systems, enabling them to see through occlusions and perceive objects at greater distances. Unlike early collaboration [2] that shares raw data and late collaboration [3, 4] that shares outputs, state-of-the-art methods are usually intermediate collaboration [5, 6] that shares features, as they can achieve the best performance-bandwidth trade-off while also exhibiting good robustness to noises.

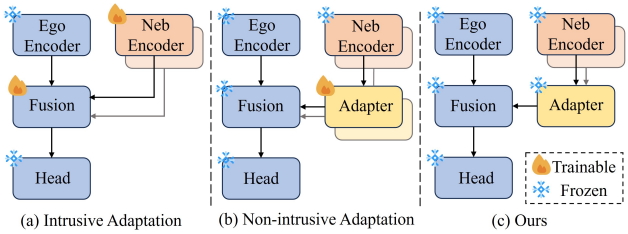


Figure 1: Frameworks for heterogeneous collaboration.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Intermediate collaborative perception typically employs homogeneous networks to achieve semantically aligned feature exchange, and numerous studies [7, 8, 9, 10, 11, 12] have been conducted to advance this field. However, in practical applications, due to the technological barriers and privacy concerns, different companies often equip their vehicles with different perception models and use different training datasets. This leads to significant domain gaps between shared features, hindering efficient and effective feature fusion between heterogeneous agents and thus degrading collaborative performance.

Collaboration between heterogeneous agents has always been a difficulty. As shown in Figure 1, existing approaches for heterogeneous collaboration can be broadly categorized into two types:

(a) Intrusive Adaptation: HEAL [13] proposes retraining the neighbor agent's encoder to align its feature to the ego-domain. HM-ViT [14] designs a heterogeneous 3D graph transformer for inter-agent and intra-agent interactions to efficiently fuse features from multi-view images and Lidar point clouds. As shown in Figure 1(a), this type of methods are intrusive, which either retraining the encoder or the fusion module. This disrupts the established original homogeneous collaboration system, making them impractical for real-world applications.

(b) Non-intrusive Adaptation: As shown in Figure 1(b), these methods [15, 16, 17, 18, 19, 20, 21] design an adapter that can be added as a plug-and-play plugin to an existing collaboration system, thus achieving adaptation without affecting the original homogeneous collaboration. However, due to the one-to-one design, it requires training a new adapter for each new heterogeneous agent, which limits their generalization and scalability. The number of adapters increases with the growth of agents, resulting in a large accumulation of adapters, which is unsustainable for resource-constrained in-vehicle devices.

To address these issues, in this paper, we explore the generalizable many-to-one domain adaptation in heterogeneous collaborative perception, that is, using only one single fixed adapter, as shown in Figure 1(c). In heterogeneous scenarios, the extracted feature should contain both domain-invariant and domain-specific information. The former represents the scene and contains inherent information about the scene, such as object spatial geometry and road topology. The latter represents the encoder and consists of specific patterns introduced by the particular encoder, thus creating domain gaps between heterogeneous agents, as discussed in MPDA [19]. Therefore, domain adaptation is essentially a process of preserving domain-invariant information while transforming domain-specific information.

Building on the above observations, we propose UniMTO, a unified many-to-one domain adaptation framework, whose core idea is to treat domain adaptation as a pattern transformation task. Specifically, unlike existing one-to-one adapters that learn rigid mappings, our framework explicitly disentangles domain-invariant content from domain-specific pattern and subsequently injects the target pattern into the source content for adaptation. This dynamic paradigm enables a single fixed adapter to universally adapt to diverse heterogeneous encoders without the need for pair-wise training. UniMTO consists of three components: the domain-invariant content encoder (DICE), responsible for extracting domain-invariant scene content; the domain-specific pattern encoder (DSPE), responsible for extracting domain-specific pattern introduced by the specific encoder; and the dynamic MoE recombiner (DMR), responsible for combining content and pattern to achieve pattern transformation. Furthermore, to facilitate robust feature separating, we propose a cross-domain content consistency loss (CCCL) to align the latent spaces of heterogeneous agents, and a content-pattern orthogonality loss (CPOL) to minimize the mutual information between content and pattern features, preventing mutual contamination and interference.

To evaluate UniMTO, we conduct extensive experiments on two real-world datasets: V2V4Real and V2X-Real. Results show that UniMTO demonstrates superior generalization, outperforming all compared baselines. It also exhibits strong zero-shot adaptation capabilities, achieving many-to-one domain adaptation using only a single, fixed adapter. Our contributions can be summarized as follows:

1) We propose UniMTO, a unified many-to-one domain adaptation framework for heterogeneous collaborative perception. By separating and recombining domain-invariant content and domain-specific pattern, pattern transformation is achieved, thus realizing domain adaptation.

2) We propose a cross-domain content consistency loss and a content-pattern orthogonality loss to promote robust feature separation, preventing mutual contamination between content and pattern.

3) Comprehensive experiments demonstrate that UniMTO achieves superior generalization, and exhibits strong zero-shot adaptation capabilities, enabling many-to-one domain adaptation with a single, fixed adapter.

2 Related Work

2.1 Homogeneous Collaborative Perception

Through information sharing, collaborative perception can solve the long-distance sparseness and occlusion problems that single-agent perception cannot overcome. Collaborative perception studies how to use the information of neighboring agents to enhance the perception ability of the ego agent itself. V2VNet [9] is a pioneering work that generate motion forecasts by sharing intermediate features. AttFuse [5] uses self-attention mechanism to fuse features while Fcooper [22] adopts the maxout operation. DiscoNet [23] achieves teacher-to-student transfer through knowledge distillation. V2X-ViT [6] uses window-based attention to fuse features of multiple agents while CoBEVT [7] designs a fused axial attention module which captures sparsely local and global spatial interactions across agents to efficiently fuse features. HM-ViT [14] combines the two to achieve multi-modal collaboration. SCOPE [24] proposes a collaborative perception framework that aggregates the spatial-temporal awareness characteristics across on-road agents. HEAL [13] and UMC [11] both propose multi-scale feature fusion methods and achieve good performance. MSD [15] proposes a pyramid deformable fusion algorithm, which is highly robust to various noises and achieves remarkable performance.

2.2 Heterogeneous Collaborative Perception

Due to the technological barriers and privacy concerns, heterogeneous collaborative perception is the true reality. Recent research has already been addressing the heterogeneity among agents. MPDA [19] proposes a learnable resizer to align features in multiple dimensions and a sparse cross-domain transformer for domain adaptation. PnPDA [17] proposes a plug-and-play adapter to address heterogeneous features. HEAL [13] proposes to sustain a unified feature space, and new agents join the collaboration through backward alignment. PLDA [15] employs an MLN-based adapter to handle domain adaptation under privacy protection. PolyInter [20] proposes a polymorphic interpreter to improve perception precision in immutable heterogeneous collaboration. GenComm [16] proposes a generation-based communication mechanism, enabling seamless perception among heterogeneous agents through feature generation. STAMP [18] proposes a scalable task-and-model agnostic collaborative perception framework, employing adapter-reverter pairs to transform features into a unified protocol domain. NegoCollab [21] introduces a negotiator to generate common representations from local representations. However, all these existing methods are designed for one-to-one scenarios, requiring retraining of the adapter whenever a completely new and unknown agent appears. Beyond this, UniMTO aims for generalization to entirely new agents, avoiding the cumbersome retraining process.

3 Method

3.1 Framework Overview

The design of UniMTO stems from a fundamental insight: despite the heterogeneity of encoders, the extracted BEV features share the same underlying scene structure, differing only in their superficial patterns. Therefore, we formulate domain adaptation as a pattern transformation task. We separate the domain-invariant content and domain-specific pattern, and generate the adapted features by injecting the ego vehicle's pattern into the collaborator's content.

For simplicity, consider a scenario with two agents, an ego vehicle and a collaborator (vehicle or infrastructure), both of which are equipped with Lidar and have the ability to perceive, communicate and detect. As shown in Figure 2, similar to standard heterogeneous collaboration, our proposed UniMTO framework consists of four stages: feature encoding, communication&alignment, domain adaptation, and fusion&detection. The process works as follows: First, in the feature encoding stage, the ego vehicle and the collaborator use their respective encoders f^e_{enc}, f^c_{enc} to extract features from the raw point clouds, $F_e = f^e_{enc}(O_e)$, $F_c = f^c_{enc}(O_c)$; Then, in the communication&alignment stage,

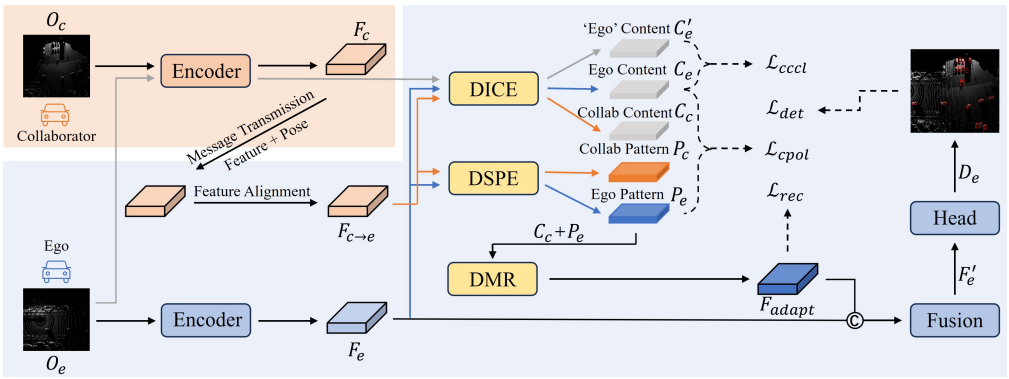


Figure 2: The overall architecture of the heterogeneous collaboration framework. The adapter consists of three parts: Domain-invariant Content Encoder (DICE), Domain-specific Pattern Encoder (DSPE) and Dynamic MoE Recombiner (DMR). $\mathcal{L}_{cccl}$, $\mathcal{L}_{cpol}$, $\mathcal{L}_{det}$ and $\mathcal{L}_{rec}$ are the training losses.

the collaborator transmits its feature F_c and pose ξ_c, and the ego vehicle performs feature alignment f_{align} based on the poses of both, $F_{c\to e} = f_{align}(F_c, \xi_c, \xi_e)$; Next, in the domain adaptation stage, the content encoder f_{ce} is used to extract the collaborator's content $C_c = f_{ce}(F_{c\to e})$, the pattern encoder f_{pe} is used to extract the ego vehicle's pattern $P_e = f_{pe}(F_e)$, and then both are input to the recombiner f_{re} to obtain the adapted feature $F_{adapt} = f_{re}(C_c, P_e)$; Finally, in the fusion&detection stage, the final detection results D_e is obtained after passing through the feature fusion module f_{fusion} and head f_{head}, $F_e^{'} = f_{fusion}(F_e, F_{adapt})$, $D_e = f_{head}(F_e^{'})$.

Unlike existing one-to-one adaptation methods that require training a specialized adapter for each heterogeneous agent, our framework only requires one adapter, meaning a single fixed adapter can handle multiple domain adaptations. The adapter includes a domain-invariant content encoder (DICE), a domain-specific pattern encoder (DSPE), and a dynamic MoE recombiner (DMR), which are introduced in the following sections, together with our designed training losses.

3.2 Model Components

Domain-invariant Content Encoder. Extracting domain-invariant content from heterogeneous features is the cornerstone of our approach. Conceptually, this content represents the inherent properties of the current scene, abstracting underlying spatial semantics such as object occupancy and road connectivity, while stripping away texture variations introduced by specific encoders.

GenComm [16] uses deformable convolution to separate foreground and background to extract spatial message, but ignores the background noise caused by feature sparsity. Therefore, as shown in Figure 3(a), we design a domain-invariant content encoder (DICE) based on deformable attention [25] to extract content, which captures global topology by dynamically sampling informative reference points, achieving both computational efficiency and robustness to domain-specific noise. It can be expressed as:

$$C(q) = \sum_{a=1}^{A} W_a \sum_{s=1}^{N_s} \phi(W_b F(q))_s F(q + \triangle q_s), \tag{1}$$

where a indexes the attention head, $W_{a/b}$ are the learnable parameters, ϕ is the softmax function. Each query q learns the 2D spatial offset $\{\triangle q_s | 1 \leq s \leq N_s\}$ where N_s is the number of sampling points. Notably, before extracting the content, we first use bilinear interpolation to align the dimensions between heterogeneous features.

Domain-specific Pattern Encoder. As discussed in MPDA [19], features extracted by different encoder have different patterns, which is the root cause of the domain gaps between heterogeneous agents. Conceptually, this pattern, as a domain identifier, encapsulates the non-geometric properties of the feature map. It reflects how the network interprets and encodes the scene, capturing the unique distribution and texture details introduced by encoder's specific operators, thus acting as the 'style' that renders the scene content.

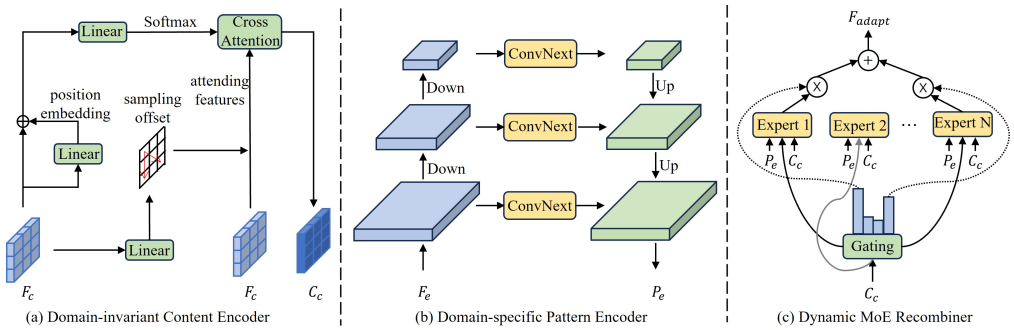


Figure 3: (a) The architecture of Domain-invariant Content Encoder (DICE). (b) The architecture of Domain-specific Pattern Encoder (DSPE). (c) The architecture of Dynamic MoE Recombiner (DMR).

Therefore, we propose a domain-specific pattern encoder (DSPE) to extract the pattern information embedded in the specific encoder. As shown in Figure 3(b), we adopt a multi-scale structure based on feature pyramid network (FPN) [26] to capture the hierarchical nature of these encoder-specific features, with a ConvNext [27] block in each scale. Finally, through top-down multi-scale integration, the coarse-grained semantics and fine-grained textures are fused to get the domain-specific pattern.

Dynamic MoE Recombiner. By reorganizing the scene-related content and the encoder-related pattern, it is theoretically possible to obtain the features extracted by a specific encoder in a specific scene. However, in practical implementation, given the vast diversity and significant discrepancies among heterogeneous encoders, relying solely on a single static network to handle such a complex many-to-one task is inherently suboptimal. It is challenging for a shared set of parameters to simultaneously accommodate the distinct distributions of all potential domains.

Therefore, drawing inspiration from the remarkable success of Mixture-of-Experts (MoE) [28] in Large Language Models (LLM) [29], which expands model capacity by conditionally activating specialized sub-networks, we design a Dynamic MoE Recombiner (DMR) to address this limitation.

Specifically, as shown in Figure 3(c), DMR consists of a gating network and several parallel experts. The gating network is a lightweight convolution block used to generate dynamic importance weights for each expert. Finally, the weighted sum of the outputs of all experts yields the adapted features, as follow:

$$F_{adapt} = \sum_{i=1}^{N} \alpha_i F_i | \sum_{i=1}^{N} \alpha_i = 1, \tag{2}$$

where N is the number of experts, and α_i and F_i are the weights and outputs of the i-th expert, respectively. All experts share an identical model architecture based on fused axial attention (FAX) transformer [7]. The domain-invariant content C serves as the query while the domain-specific pattern P serves as the key and value, allowing each expert to adaptively inject pattern information into the inherent scene. The whole process can be formulated as below:

$$Q = W_Q C, K = W_K P, V = W_V P, \tag{3}$$

$$F^{'} = Q + LN(FAX(Q, K, V)), \tag{4}$$

$$F^{''} = F^{'} + LN(FFN(F^{'})), \tag{5}$$

where LN is layer normalization, FFN is feed forward network, FAX is fused axial attention, Q is query, K is key, and V is value, respectively. W_Q, W_K and W_V are learnable parameters.

3.3 Loss Functions

Cross-domain Content Consistency Loss. Since the content encoder extracts inherent information from the scene, for the same point cloud input, the extracted content should be consistent regardless of which encoder is used. Thus we propose cross-domain content consistency loss (CCCL). Specifically, as shown in Figure 1, we feed the ego vehicle's point cloud simultaneously to the encoders of both

ego vehicle and collaborator, and the subsequent DICE block to obtain Ego Content C_e and 'Ego' Content $C_e^{'}$, respectively, minimizing the distance between them:

$$\mathcal{L}_{cccl} = ||C_e - C_e^{'}||_2^2. \tag{6}$$

By minimizing $\mathcal{L}_{cccl}$, we force the DICE to extract the inherent elements within the scene.

Content-Pattern Orthogonality Loss. To ensure robust feature separation, it is critical to prevent information leakage between content and pattern. Ideally, they should be independent of each other. To achieve this, we propose content-pattern orthogonality loss (CPOL) to minimize the mutual information. Specifically, we compute the square of the cosine similarity between the global vectors of content and pattern:

$$\mathcal{L}_{cpol} = (\frac{gap(C_e) \cdot gap(P_e)}{||gap(C_e)||_2 \cdot ||gap(P_e)||_2})^2 + (\frac{gap(C_c) \cdot gap(P_c)}{||gap(C_c)||_2 \cdot ||gap(P_c)||_2})^2, \tag{7}$$

where gap is the global average pooling, C_e, P_e, C_c, P_c are Ego Content, Ego Pattern, Collab Content, and Collab Pattern, respectively, as shown in Figure 2. By minimizing $\mathcal{L}_{cpol}$, we force the content and pattern to be orthogonal in the latent space, thereby minimizing their correlation and ensuring the purity of the separation.

Total Loss. In addition to the losses mentioned above, we incorporate reconstruction loss and detection loss to ensure the fidelity of feature domain adaptation and downstream task performance, as below:

$$\mathcal{L}_{rec} = mse(F_{adapt}, f_{align}(F_{gt})), \tag{8}$$

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg}, \tag{9}$$

where F_{adapt} is the output of recombiner, F_{gt} is the ground truth, i.e., the feature extracted from the collaborator's point cloud by the ego vehicle's encoder, $\mathcal{L}_{cls}$ is the classification loss, and $\mathcal{L}_{reg}$ is the regression loss. The total training loss is formulated as:

$$\mathcal{L}_{total} = \lambda_{det}\mathcal{L}_{det} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cccl}\mathcal{L}_{cccl} + \lambda_{cpol}\mathcal{L}_{cpol}, \tag{10}$$

where λ_{det}, λ_{rec}, λ_{cccl}, and λ_{cpol} are hyperparameters balancing the contribution of different terms.

4 Experiments

4.1 Settings

Datasets and Evaluation Metrics. We conduct experiments on two datasets: V2V4Real [30] and V2X-Real [31]. V2V4Real is the first large-scale real-world dataset for V2V perception in autonomous driving. It is collected by two vehicles simultaneously in the same scene, providing multi-view multi-sensor datastream, covering various road sections including city roads, highways and freeways. V2X-Real is also a real-world dataset but on a larger scale, consisting two vehicles and two infrastructure, which are all equipped with multi-modal sensors including Lidar and cameras. We only use the car category of objects. Our experiments only focus on Lidar-based 3D object detection, the detection performance is measured with Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7.

Heterogeneous Agents Setup. PointPillar [32], SECOND [33], and VoxelNet [34] are commonly used Lidar feature encoders, so we set up six heterogeneous encoders with different configurations, as shown in Table 1. We train all six encoders on both V2V4Real and V2X-Real, resulting in a total of twelve heterogeneous agents. When evaluating the adapter, we consider the following three heterogeneous scenarios according to the collaborator's encoder:

1) Hetero Scenario 1: Training with sd1 and inference with sd1 to verify the one-to-one adaptation capabilities.

2) Hetero Scenario 2: Training with sd1, vn8, pp4 and inference with sd1, vn8, pp4 to verify the many-to-one adaptation capabilities and generalization.

3) Hetero Scenario 3: Training with sd1, vn8, pp4 but inference with sd2, vn4, pp8 to verify the zero-shot adaptation capabilities.

In all experiments, the ego vehicle's encoder is pp4-ego, with a fusion module of AttFuse [5] or V2X-ViT [6]. In scenarios 2 and 3, during training, we randomly sample from sd1, vn8, and pp4 in each iteration as the collaborator's encoder. Note that all encoder, fusion module and detection head are pre-trained and frozen, only the adapter is trainable.

Table 1: Detailed configurations of heterogeneous encoders. pp4 and pp8 are two versions of PointPillar, sd1 and sd2 are versions of SECOND, vn4 and vn8 are versions of VoxelNet.

Encoder	Voxel Size (x, y)	Half Lidar Range (x, y)	Feature Shape (C×H×W)
pp4	0.4, 0.4	102.4, 38.4	256×96×256
pp8	0.8, 0.8	102.4, 38.4	256×48×128
sd1	0.1, 0.1	102.4, 38.4	256×96×256
sd2	0.2, 0.2	102.4, 38.4	256×48×128
vn4	0.4, 0.4	102.4, 38.4	256×96×256
vn8	0.8, 0.8	102.4, 38.4	256×48×128

Table 2: One-to-one adaptation performance in Hetero Scenario 1. The fusion method is V2X-ViT.

Method	V2V4Real		V2X-Real	
	AP0.5	AP0.7	AP0.5	AP0.7
No Fusion	56.0	38.6	58.6	49.9
NegoCollab[21]	<u>70.2</u>	45.6	73.1	<u>59.1</u>
PLDA[15]	69.2	45.1	<u>73.5</u>	<u>59.1</u>
PnPDA[17]	68.0	42.3	68.9	33.5
GenComm[16]	69.0	<u>45.8</u>	71.3	57.8
STAMP[18]	69.2	44.2	72.8	58.6
MPDA[19]	69.9	44.0	72.4	58.4
PolyInter[20]	65.1	43.3	70.0	56.4
UniMTO (Ours)	**70.7**	**46.9**	**73.7**	**59.5**

Implementation Details. The communication distance for V2V4Real and V2X-Real are set to 70m and 50m, respectively. The evaluation range in x and y directions are [-102.4m, 102.4m] and [-38.4m, 38.4m], respectively. All models are trained end-to-end for 60 epochs on RTX 3090 24GB. Adam [35] is used for optimization. An initial learning rate of 0.001 is selected, and is multiplied by 0.1 every 20 epochs during training. Early stop is used to find the best epoch. In all experiments, λ_{det}, λ_{rec}, λ_{cccl}, λ_{cpol}, N_s and N are set to 1, 1, 2, 10, 9 and 3, respectively.

4.2 Quantitative Results

One-to-One Adaptation. Table 2 compares the one-to-one adaptation performance of the proposed UniMTO with seven state-of-the-art methods, NegoCollab, PLDA, PnPDA, GenComm, STAMP, MPDA, PolyInter, and one typical baseline No Fusion, which means single-agent detection without any collaboration. It can be seen that all methods achieve comparable performance, significantly outperforming No Fusion, which demonstrates how a plug-and-play adapter can bridge the domain gap through dedicated training. Furthermore, among all these methods, UniMTO performs best, specifically, on V2V4Real, the AP0.5/AP0.7 is at least 0.5%/1.1% higher than that of others, while it is 0.2%/0.4% on V2X-Real. Its AP0.5/AP0.7 are 14.7%/8.3% and 15.1%/9.6% higher than No Fusion on both datasets, respectively, showcasing its strong one-to-one domain adaptation capability.

Many-to-One Adaptation on Seen Domains. Table 4 compares the many-to-one adaptation performance, where Homo Collab represents homogeneous collaboration, i.e., all agents share the same encoder. It can be seen that compared to the one-to-one dedicated training, the performance of all methods decreases in the many-to-one scenario. Among these, UniMTO remains the SOTA, and surprisingly, it not only significantly outperforms No Fusion but is also the only one whose performance is comparable to or even surpasses Homo Collab in all experiments. Specifically, on V2X-Real, when the collaborator's encoder is sd1/vn8/pp4, UniMTO's AP0.5 exceeds No Fusion 12.7%/12.8%/12.4%, and exceeds Homo Collab 2.2%/2.3%/1.9%, respectively. Results demonstrate UniMTO's strong generalization ability, indicating that through content-pattern decoupling and recombination, a single adapter can adapt to multiple encoders.

Zero-shot Adaptation on Unseen Domains. In Table 5, we verify a scenario where the encoder during inference is unseen during training, which more closely reflects real-world situations. It can be seen that under such stringent conditions, the performance of other methods is significantly lower than Homo Collab, and PnPDA even performs worse than No Fusion. In contrast, UniMTO demonstrates strong zero-shot adaptation capabilities, significantly outperforming No Fusion and closely approaching Homo Collab. Specifically, on V2X-Real, when the collaborator's encoder is sd2/vn4/pp8, its AP0.5 is only 0.1%/0.3%/0.1% lower than Homo Collab, and is at least 2.3%/2.0%/2.3% higher than the second-best method.

Table 4: Many-to-one adaptation performance on seen domains in Hetero Scenario 2. The bold ones are the highest, and the underlined ones are the second. Homo Collab refers to homogeneous collaboration, where the ego vehicle and the collaborator use exactly the same encoder.

	V2V4Real						V2X-Real		
	AttFuse Fusion			V2X-ViT Fusion			V2X-ViT Fusion		
Method	sd1	vn8	pp4	sd1	vn8	pp4	sd1	vn8	pp4
No Fusion	56.6/37.0	56.6/37.0	56.6/37.0	56.0/38.6	56.0/38.6	56.0/38.6	58.6/49.9	58.6/49.9	58.6/49.9
Homo Collab	70.6/44.9	70.6/44.9	70.6/44.9	69.2/44.4	69.2/44.4	69.2/44.4	69.1/56.4	69.1/56.4	69.1/56.4
NegoCollab[21]	69.7/43.5	69.7/44.0	70.2/43.0	**68.7**/44.8	68.3/43.5	68.2/**45.3**	70.3/56.8	71.3/56.8	69.8/56.9
PLDA[15]	56.6/37.1	58.5/39.2	56.6/37.1	58.3/40.0	57.3/38.4	59.2/40.8	60.1/51.8	61.7/53.0	62.0/53.2
PnPDA[17]	59.8/11.1	60.8/12.6	59.5/12.4	64.8/26.1	63.6/24.7	63.8/23.6	67.0/45.8	66.8/45.8	67.0/45.9
GenComm[16]	69.0/43.8	69.9/43.0	70.1/44.0	67.6/44.4	68.2/43.8	67.4/44.0	70.9/56.3	70.0/56.2	71.0/56.4
STAMP[18]	70.3/**45.0**	**71.7**/42.6	70.3/43.7	68.0/43.2	66.4/43.6	68.3/44.3	69.7/55.3	69.9/56.8	70.2/57.1
MPDA[19]	67.7/38.9	65.7/37.8	68.1/39.1	67.5/43.4	67.8/44.2	68.1/43.0	70.3/57.1	70.4/**57.6**	70.1/57.0
PolyInter[20]	67.6/42.4	67.6/41.8	67.7/41.5	64.0/42.2	64.2/42.8	64.3/43.4	66.8/54.9	63.8/53.1	66.9/54.8
UniMTO (Ours)	**70.4**/44.5	70.7/**44.6**	**70.7/44.5**	68.5/**45.1**	**68.6/45.5**	**68.9**/45.1	**71.3/57.4**	**71.4**/57.5	**71.0/57.1**

Table 5: Zero-shot adaptation performance on unseen domains in Hetero Scenario 3. The bold ones are the highest, and the underlined ones are the second. Homo Collab refers to homogeneous collaboration, where the ego vehicle and the collaborator use exactly the same encoder.

	V2V4Real						V2X-Real		
	AttFuse Fusion			V2X-ViT Fusion			V2X-ViT Fusion		
Method	sd2	vn4	pp8	sd2	vn4	pp8	sd2	vn4	pp8
No Fusion	56.6/37.0	56.6/37.0	56.6/37.0	56.0/38.6	56.0/38.6	56.0/38.6	58.6/49.9	58.6/49.9	58.6/49.9
Homo Collab	70.6/44.9	70.6/44.9	70.6/44.9	69.2/44.4	69.2/44.4	69.2/44.4	69.1/56.4	69.1/56.4	69.1/56.4
NegoCollab[21]	57.2/37.7	55.4/36.8	45.7/32.3	54.2/40.3	55.8/35.2	55.8/35.2	61.2/50.2	61.4/52.0	58.9/51.2
PLDA[15]	56.6/37.1	56.6/37.0	56.6/37.1	54.9/39.2	55.4/39.7	53.1/36.2	60.3/52.0	57.6/50.1	56.3/48.9
PnPDA[17]	31.6/5.8	42.3/6.3	17.1/1.7	47.4/13.1	44.6/15.4	27.8/11.1	66.7/46.1	66.8/45.8	66.7/45.8
GenComm[16]	48.1/32.1	54.9/36.8	56.6/37.0	54.4/40.4	55.5/40.5	55.3/40.0	59.9/50.6	60.2/51.1	59.4/50.4
STAMP[18]	48.1/31.3	53.4/36.0	55.5/36.3	54.4/38.8	54.2/39.3	48.1/35.5	61.6/52.1	64.3/54.7	54.3/46.4
MPDA[19]	59.5/37.2	53.2/35.8	57.6/36.5	56.2/40.9	58.3/41.4	57.6/41.6	62.6/51.3	62.9/52.5	47.9/42.4
PolyInter[20]	56.4/35.8	55.0/35.6	56.7/36.5	52.8/38.0	57.4/39.2	55.7/39.0	62.8/52.4	63.3/52.7	58.2/49.7
UniMTO (Ours)	**63.3/40.9**	**59.7/38.5**	**68.1/42.6**	**62.0/44.0**	**61.4/42.9**	**66.7/44.7**	**69.0/56.2**	**68.8/55.9**	**69.0/56.5**

The experimental results verify the feasibility of the framework we proposed. When collaborating participants are equipped with different encoders due to practical considerations, a single adapter can handle all encoders. This not only solves the domain adaptation problem but also avoids the unsustainable issues faced by existing one-to-one approaches due to limited in-vehicle resources. This framework is more in line with actual deployment and provides a new perspective for the study of heterogeneous collaborative perception. Experiments with more fusion methods, experiments on cross-dataset and cross-modal adaptation, experiments about simultaneous collaboration and leave-one-family-out evaluation, computational overhead analysis are presented in Appendix A, B, C, D, E, and F.

Table 3: Ablation results in Hetero Scenario 3 on V2V4Real. The collaborator's encoder is sd2. The fusion method is AttFuse.

DICE	DSPE	DMR	CCCL	CPOL	AP0.5	AP0.7
					56.6	37.0
✓	✓	✓			57.3	37.1
✓	✓	✓	✓		61.5	40.4
✓	✓	✓		✓	58.8	37.4
✓	✓	✓	✓	✓	**63.3**	**40.9**
	✓	✓	✓	✓	62.0	40.4
✓			✓	✓	56.9	37.8

4.3 Ablation Study

Effect of Components in UniMTO. We conduct ablation studies to analyze the effectiveness of the proposed components. We use No Fusion as our baseline and progressively add components to evaluate the contribution of each component. As shown in Table 3, all components contribute

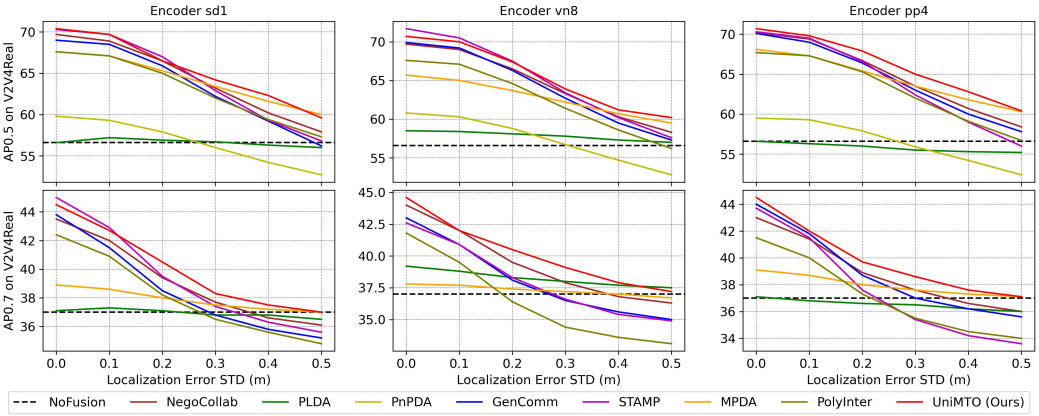


Figure 4: Robustness to localization error in Hetero Scenario 2 on V2V4Real, the fusion method is AttFuse.

positively to performance improvement. Specifically, the absence of DICE and DSPE&DMR reduces AP0.7 by 0.5% and 3.1%, while the absence of CCCL and CPOL reduces it by 3.5% and 0.5%, respectively.

Ablations of DMR Design. We conduct additional ablations, varying the number of experts, gating strategy, and recombination method. As shown in Table 6, performance nearly saturates with three experts, we therefore use three experts as a favorable performance-complexity trade-off. Dynamic softmax gating achieves the best performance, while uniform averaging and static learned gating provide limited gains despite comparable model capacity, which demonstrates the effectiveness of dynamic MoE and shows that the improvement cannot be explained by increased model capacity alone. Dynamic MoE recombination also outperforms the Add+Conv and Concat+Conv alternatives. Together, these ablations further demonstrate the necessity of DMR.

Table 6: Ablations on the number of experts, gating strategy, and recombination method in Hetero Scenario 2 on V2V4Real. The collaborator's encoder is sd1.

Configuration	AP0.5	AP0.7
Impact of expert number N		
N=1	68.4	42.8
N=2	69.6	43.2
N=3	70.4	44.5
N=4	70.6	44.4
N=5	70.5	44.8
Gating strategy comparison		
Uniform averaging	69.3	42.6
Static learned gate	69.8	43.4
Dynamic softmax gate (Ours)	**70.4**	**44.5**
Recombination method comparison		
Add+Conv	56.7	37.1
Concat+Conv	66.2	39.1
Dynamic MoE (Ours)	**70.4**	**44.5**

Robustness to Localization Errors. We verify the performance of UniMTO at different levels of localization error on both V2V4Real and V2X-Real datasets. Specifically, we add Gaussian noise with a variance of 0 and a mean ranging from 0 to 0.5 meter to the collaborator's pose. Figure 4 shows that the performance of all models decreased as the pose error increased, due to feature misalignment. When the error reaches a large value, the AP0.5 of PLDA, PnPDA are even lower than the No Fusion baseline. Noticeably, UniMTO outperforms all other methods, and No Fusion at most error level, which demonstrates the robustness of UniMTO against collaboration pose noises. Detailed values of Figure 4 are available in Appendix H.

Computational Overhead and Performance. We evaluate UniMTO and all compared methods under the same inference setting on V2V4Real and report adapter parameters, peak GPU memory, end-to-end inference latency, adapter FLOPs, and detection performance. Specifically, parameters and FLOPs count the adapter only, peak memory measures end-to-end inference, and latency includes preprocessing, forward, and postprocessing. As shown in Table 7, UniMTO achieves the best detection performance while maintaining computational overhead comparable to existing methods, with only one shared adapter. Specifically, compared with the second-best MPDA, UniMTO adds

Table 7: Computational overhead and performance comparison in Hetero Scenario 3 on V2V4Real. The collaborator's encoder is sd2. The fusion method is AttFuse.

Method	Params (M)	Peak memory (MB)	Latency (ms)	FLOPs (G)	AP0.5	AP0.7
NegoCollab[21]	5.6	870.5	156.4	154.3	57.2	<u>37.7</u>
PLDA[15]	11.5	950.3	135.5	282.2	56.6	37.1
PnPDA[17]	2.7	1678.5	157.2	66.9	31.6	5.8
GenComm[16]	2.5	1184.0	158.2	30.8	48.1	32.1
STAMP[18]	2.3	840.1	133.8	42.9	48.1	31.3
MPDA[19]	6.2	1041.8	149.6	142.5	<u>59.5</u>	37.2
PolyInter[20]	28.2	1089.2	123.6	84.1	56.4	35.8
UniMTO (Ours)	9.8	1537.5	155.0	117.8	**63.3**	**40.9**

3.6M adapter parameters, 495.7MB peak memory, and 5.4ms latency, while requiring 24.7G fewer adapter FLOPs.

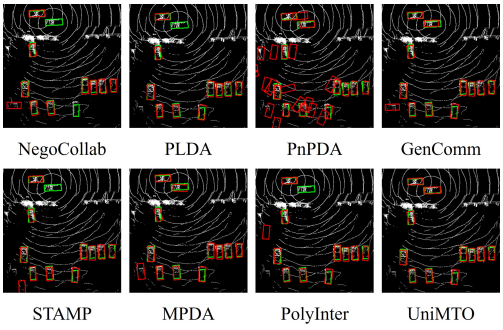


Figure 5: 3D detection visualization on V2V4Real in Hetero Scenario 3.

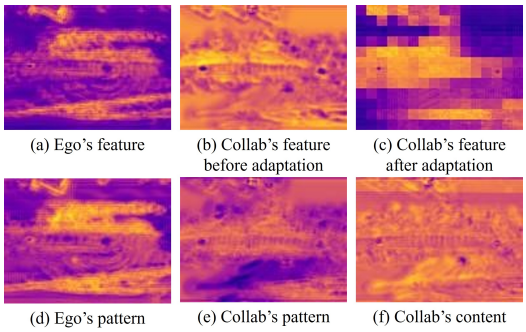


Figure 6: Visualization results on V2V4Real in Hetero Scenario 2.

4.4 Qualitative Analysis

We conduct a qualitative analysis by visualizing typical samples from the V2V4Real dataset. Figure 5 shows the detection results of UniMTO and all compared methods. It can be seen that UniMTO predicts more accurate bounding boxes while other approaches exhibit larger displacements. Furthermore, to provide a more intuitive demonstration of UniMTO's effectiveness, we sum the features along the channel dimension, with visualizations shown in Figure 6. It can be seen that there is a significant domain gap between the original features of the ego vehicle and the collaborator, but after UniMTO's adaptation, they are visually much closer. Simultaneously, the content and pattern display clearly distinguishable distributions, validating that the model learns to disentangle different aspects of the input data. More visualizations are available in Appendix I.

5 Conclusion

We propose UniMTO, a unified domain adaptation framework, which enables many-to-one domain adaptation using a single, fixed adapter, saving the significant pairwise training costs and reducing the deployment pressure for resource-constrained in-vehicle devices. Extensive experiments covering multiple scenarios demonstrate its SOTA performance. We hope that our work can accelerate the actual implementation of collaborative perception.

6 Acknowledgements

This work is funded by National Key Laboratory of Collective Intelligence & Collaboration (Open Fund Project No. QXZ25017101), and by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grant No. 2025C01039 and 2025C02164, and by the Key Project of Natural Science Foundation of Zhejiang Province under Grant LZ26F010003.

References

[1] Si Liu, Chen Gao, Yuan Chen, Xingyu Peng, Xianghao Kong, Kun Wang, Runsheng Xu, Wentao Jiang, Hao Xiang, Jiaqi Ma, et al. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. *arXiv preprint arXiv:2308.16714*, 2023.

[2] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019.

[3] Zaydoun Yahya Rawashdeh and Zheng Wang. Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3961–3966. IEEE, 2018.

[4] Runsheng Xu, Weizhe Chen, Hao Xiang, Xin Xia, Lantao Liu, and Jiaqi Ma. Model-agnostic multi-agent perception framework. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1471–1478. IEEE, 2023.

[5] 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.

[6] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022.

[7] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022.

[8] 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.

[9] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 605–621. Springer, 2020.

[10] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

[11] Tianhang Wang, Guang Chen, Kai Chen, Zhengfa Liu, Bo Zhang, Alois Knoll, and Changjun Jiang. Umc: A unified bandwidth-efficient and multi-resolution based collaborative perception framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8187–8196, 2023.

[12] 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, 2024.

[13] Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Yanfeng Wang, and Siheng Chen. An extensible framework for open heterogeneous collaborative perception. *arXiv preprint arXiv:2401.13964*, 2024.

[14] Hao Xiang, Runsheng Xu, and Jiaqi Ma. Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 284–295, 2023.

[15] Bin Lu, Xinyu Xiao, Changzhou Zhang, Yang Zhou, Zhiyu Xiang, Hangguan Shan, and Eryun Liu. Privacy-preserving v2x collaborative perception integrating unknown collaborators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5802–5810, 2025.

[16] Junfei Zhou, Penglin Dai, Quanmin Wei, Bingyi Liu, Xiao Wu, and Jianping Wang. Pragmatic heterogeneous collaborative perception via generative communication mechanism. *arXiv preprint arXiv:2510.19618*, 2025.

[17] Tianyou Luo, Quan Yuan, Guiyang Luo, Yuchen Xia, Yujia Yang, and Jinglin Li. Plug and play: A representation enhanced domain adapter for collaborative perception. In *European Conference on Computer Vision*, pages 287–303. Springer, 2024.

[18] Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. Stamp: Scalable task and model-agnostic collaborative perception. *arXiv preprint arXiv:2501.18616*, 2025.

[19] Runsheng Xu, Jinlong Li, Xiaoyu Dong, Hongkai Yu, and Jiaqi Ma. Bridging the domain gap for multi-agent perception. *arXiv preprint arXiv:2210.08451*, 2022.

[20] Yuchen Xia, Quan Yuan, Guiyang Luo, Xiaoyuan Fu, Yang Li, Xuanhan Zhu, Tianyou Luo, Siheng Chen, and Jinglin Li. One is plenty: A polymorphic feature interpreter for immutable heterogeneous collaborative perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1592–1601, 2025.

[21] Shao Congzhang, Quan Yuan, Guiyang Luo, Yue Hu, Danni Wang, Liu Yilin, Rui Pan, Bo Chen, and Jinglin Li. Negocollab: A common representation negotiation approach for heterogeneous collaborative perception. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[22] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pages 88–100, 2019.

[23] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021.

[24] Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, and Liang Song. Spatio-temporal domain awareness for multi-agent collaborative perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23383–23392, 2023.

[25] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[28] Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

[29] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[30] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023.

[31] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, et al. V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision*, pages 455–470. Springer, 2024.

[32] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

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

[34] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022.

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[38] Qi Zhang. A novel resnet101 model based on dense dilated convolution for image classification. *SN Applied Sciences*, 4(1):9, 2022.

[39] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020.

A Experiments with More Fusion Methods

We further use V2VNet [9] as the fusion method and conduct experiments on all scenarios in Hetero Scenario 1, Hetero Scenario 2 and Hetero Scenario 3. The experimental results are shown in Table 8, Table 9, and Table 10, respectively.

Table 8: One-to-one adaptation performance in Hetero Scenario 1.

Method	V2V4Real		V2X-Real	
	AP0.5	AP0.7	AP0.5	AP0.7
No Fusion	50.1	29.5	59.9	45.8
NegoCollab[21]	69.7	38.5	77.2	59.9
PLDA[15]	70.0	40.1	76.3	59.2
PnPDA[17]	65.4	36.9	65.0	31.6
GenComm[16]	69.0	40.1	77.7	58.4
STAMP[18]	69.2	39.2	76.8	59.3
MPDA[19]	69.8	39.6	74.7	56.9
PolyInter[20]	65.6	36.3	73.7	56.3
UniMTO (Ours)	**70.7**	**40.4**	**77.8**	**60.2**

Table 9: Many-to-one adaptation performance on seen domains in Hetero Scenario 2.

Method	V2V4Real			V2X-Real		
	sd1	vn8	pp4	sd1	vn8	pp4
No Fusion	50.1/29.5	50.1/29.5	50.1/29.5	59.9/45.8	59.9/45.8	59.9/45.8
Homo Collab	67.1/37.1	67.1/37.1	67.1/37.1	74.5/55.9	74.5/55.9	74.5/55.9
NegoCollab[21]	68.6/38.4	68.3/38.3	68.9/38.2	75.6/**56.8**	75.2/56.3	74.5/56.0
PLDA[15]	50.7/28.7	44.3/27.1	51.6/29.3	63.8/49.0	66.4/51.2	62.0/47.7
PnPDA[17]	65.5/32.9	63.9/30.9	65.2/31.4	59.7/30.8	60.8/31.3	60.0/30.9
GenComm[16]	68.9/38.5	68.5/37.2	68.7/**39.8**	73.8/55.6	75.0/**56.9**	74.1/56.3
STAMP[18]	69.5/38.6	**69.1**/38.1	69.1/38.2	74.3/54.7	75.2/56.2	74.4/55.5
MPDA[19]	67.4/39.0	66.3/**39.4**	67.5/39.3	74.9/56.2	74.6/56.5	73.3/54.8
PolyInter[20]	62.4/36.0	63.0/35.4	63.0/39.2	70.5/53.1	68.7/51.4	69.5/52.1
UniMTO (Ours)	**69.8/39.3**	68.9/38.4	**69.8**/39.3	**75.8**/56.3	**75.7**/56.7	**75.6/56.5**

Table 10: Zero-shot adaptation performance on unseen domains in Hetero Scenario 3.

Method	V2V4Real			V2X-Real		
	sd2	vn4	pp8	sd2	vn4	pp8
No Fusion	50.1/29.5	50.1/29.5	50.1/29.5	59.9/45.8	59.9/45.8	59.9/45.8
Homo Collab	67.1/37.1	67.1/37.1	67.1/37.1	74.5/55.9	74.5/55.9	74.5/55.9
NegoCollab[21]	51.4/27.1	44.2/25.1	51.3/21.7	63.9/47.5	62.7/47.8	52.5/39.4
PLDA[15]	48.0/27.7	46.6/28.4	49.2/29.4	62.4/47.1	59.0/45.6	63.0/47.2
PnPDA[17]	41.5/15.6	32.2/13.5	34.4/14.6	57.5/28.4	59.3/29.8	58.6/29.6
GenComm[16]	48.8/30.2	50.1/32.3	50.4/30.5	63.6/47.2	62.2/47.3	62.6/47.5
STAMP[18]	50.8/26.7	50.6/30.1	52.7/27.8	56.7/35.8	39.1/24.7	52.9/33.1
MPDA[19]	31.9/20.7	40.8/18.4	35.4/16.4	56.3/32.5	60.3/42.3	43.8/32.7
PolyInter[20]	51.8/30.8	52.4/31.1	54.7/33.1	64.5/48.4	64.0/48.4	65.7/49.4
UniMTO (Ours)	**59.0/32.6**	**60.2/34.1**	**60.1/33.9**	**70.5/51.6**	**69.6/53.7**	**70.7/52.8**

As can be seen, UniMTO achieves SOTA performance in all aspects, including one-to-one adaptation, many-to-one adaptation, and zero-shot adaptation. Specifically, in Hetero Scenario 3 on V2X-Real, when the collaborator's encoder is sd2/vn4/pp8, UniMTO's AP0.5 exceeds No Fusion by 10.6%/9.7%/10.8%, and exceeds the second-best by 6.0%/5.6%/5.0%, respectively. The experimental results further demonstrate the effectiveness and efficiency of UniMTO. It can achieve multiple domain adaptations using only a fixed adapter, saving the significant training costs of one-to-one pairwise training and reducing the deployment pressure on resource-constrained in-vehicle devices.

B Cross-dataset Adaptation

PLDA [15] mentions that, in addition to the differences in the encoder structure itself, differences in training datasets can also cause domain gaps between features. Therefore, we use the DAIR-V2X [36] dataset to verify UniMTO's cross-dataset domain adaptation capabilities. DAIR-V2X is the first large-scale, multi-modality, multi-view dataset from real scenarios for vehicle-infrastructure cooperative autonomous driving. It has 9000 frames featuring one vehicle and one road side unit, both equipped with a Lidar and a camera. In our experiments, we merge the four categories of car, van, truck and bus into one category.

Specifically, we consider the following two additional heterogeneous scenarios:

- Hetero Scenario 4: Training with sd1-dair, vn8-dair, pp4-dair and inference with sd1-dair, vn8-dair, pp4-dair to verify the many-to-one adaptation capability.
- Hetero Scenario 5: Training with sd1-dair, vn8-dair, pp4-dair and inference with sd2-dair, vn4-dair, pp8-dair to verify the zero-shot adaptation capability.

Here, -dair indicate that the collaborator's encoder are trained on the DAIR-V2X dataset.

Table 11: Many-to-one adaptation performance on seen domains in Hetero Scenario 4. The fusion method is V2X-ViT.

Method	V2V4Real			V2X-Real		
	sd1-dair	vn8-dair	pp4-dair	sd1-dair	vn8-dair	pp4-dair
No Fusion	56.0/38.6	56.0/38.6	56.0/38.6	58.6/49.9	58.6/49.9	58.6/49.9
Homo Collab	69.2/44.4	69.2/44.4	69.2/44.4	69.1/56.4	69.1/56.4	69.1/56.4
NegoCollab[21]	67.3/39.9	68.7/42.2	**69.8**/41.5	66.5/55.4	67.1/55.0	67.0/55.3
PLDA[15]	59.2/40.2	60.9/42.4	58.6/40.7	61.9/52.7	61.4/53.0	59.5/51.8
PnPDA[17]	62.0/22.0	61.6/24.3	64.0/25.0	66.7/45.7	66.8/45.8	66.6/45.7
GenComm[16]	67.4/44.0	68.7/**45.0**	68.7/44.9	65.6/55.1	68.2/56.1	68.1/56.0
STAMP[18]	68.1/43.2	69.2/43.9	69.1/44.8	66.9/55.1	68.2/56.1	67.9/55.9
MPDA[19]	64.8/40.1	**69.5**/43.0	68.9/41.1	66.7/55.6	67.7/55.8	67.5/56.1
PolyInter[20]	62.9/40.1	62.4/42.7	63.7/41.3	62.7/52.5	63.1/52.8	64.4/53.7
UniMTO (Ours)	**68.7/45.1**	68.9/44.7	69.0/**45.3**	**70.0/56.7**	**70.5/56.9**	**70.5/56.8**

Table 12: Zero-shot adaptation performance on unseen domains in Hetero Scenario 5. The fusion method is V2X-ViT.

Method	V2V4Real			V2X-Real		
	sd2-dair	vn4-dair	pp8-dair	sd2-dair	vn4-dair	pp8-dair
No Fusion	56.0/38.6	56.0/38.6	56.0/38.6	58.6/49.9	58.6/49.9	58.6/49.9
Homo Collab	69.2/44.4	69.2/44.4	69.2/44.4	69.1/56.4	69.1/56.4	69.1/56.4
NegoCollab[21]	56.0/39.0	55.6/35.9	55.4/36.1	62.9/53.0	61.9/52.1	64.2/53.2
PLDA[15]	56.6/38.6	52.2/35.4	56.9/40.4	56.8/49.1	50.6/45.2	56.4/49.0
PnPDA[17]	45.3/10.4	26.9/7.2	43.2/9.7	66.7/45.8	**66.8**/45.8	66.7/45.8
GenComm[16]	57.1/40.7	52.0/37.4	55.2/39.9	60.5/51.3	60.5/51.1	61.8/52.4
STAMP[18]	52.0/38.2	51.3/37.2	54.6/39.1	52.0/41.8	56.7/46.2	61.3/50.9
MPDA[19]	63.4/42.6	55.3/40.0	56.2/40.4	63.7/53.4	46.9/40.9	63.5/52.7
PolyInter[20]	57.8/39.4	50.6/38.6	53.3/37.4	63.4/53.1	62.7/52.2	61.9/51.5
UniMTO (Ours)	**67.8/45.2**	**66.8/45.8**	**64.1/43.8**	**68.7/56.0**	64.3/**53.0**	**68.6/55.7**

As can be seen from Table 11 and Table 12, in the cross-dataset training scenario, UniMTO demonstrates SOTA performance, regardless of whether the encoders (sd1-dair, vn8-dair, pp4-dair) were seen during training or not (sd2-dair, vn4-dair, pp8-dair). Specifically, on V2X-Real, when the collaborator's encoder is sd1-dair/vn8-dair/pp4-dair, UniMTO's AP0.5 exceeds No Fusion by 11.4%/11.9%/11.9%, exceeds Homo Collab by 0.9%/1.4%/1.4%, and exceeds the second-best method by 3.1%/2.3%/2.4%, respectively. The experimental results demonstrate UniMTO's excellent cross-dataset domain adaptation capabilities, further confirming its enormous potential for practical applications.

C Cross-modal Adaptation

Images, as data with rich semantic details, have received widespread attention and application. Therefore, we additionally set up two popular image encoders, EfficientNet [37] (eff) and ResNet101 [38] (res101), to verify UniMTO's cross-modal domain adaptation capabilities, in which we use LSS (Lift-Splat-Shot) [39] to transform front view image features into BEV features. We use the V2XSet [6] dataset, a large-scale dataset for V2X perception, where each agent is equipped with one Lidar and four cameras. In total, there are 11447 frames and the train/validate/test splits are 6694/1920/2833, respectively.

Specifically, we consider the following three additional heterogeneous scenarios:

- Hetero Scenario 6: Training with eff and inference with eff to verify the one-to-one adaptation capability.
- Hetero Scenario 7: Training with eff, res101 and inference with eff, res101 to verify the many-to-one adaptation capability.
- Hetero Scenario 8: Training with eff but inference with res101 to verify the zero-shot adaptation capability.

Table 13: Adaptation performance on Hetero Scenario 6, Hetero Scenario 7, and Hetero Scenario 8. The fusion method is V2X-ViT.

Method	Hetero Scenario 6 eff	Hetero Scenario 7 eff	Hetero Scenario 7 res101	Hetero Scenario 8 res101
No Fusion	90.6/80.9	90.6/80.9	90.6/80.9	90.6/80.9
Homo Collab	92.9/84.7	92.9/84.7	92.9/84.7	92.9/84.7
NegoCollab[21]	91.5/82.2	91.2/82.0	91.6/82.0	58.6/52.9
PLDA[15]	91.1/82.3	91.3/82.3	91.1/82.4	90.0/80.9
PnPDA[17]	74.8/52.0	74.4/37.8	74.5/37.8	71.4/50.2
GenComm[16]	90.0/81.9	90.9/81.6	90.8/81.7	90.2/81.3
STAMP[18]	91.3/81.6	91.0/81.9	91.4/82.0	90.2/80.7
MPDA[19]	90.9/81.4	90.8/81.6	91.4/81.9	84.3/76.0
PolyInter[20]	90.5/80.7	91.3/82.5	91.6/**82.9**	86.7/77.9
UniMTO (Ours)	**91.7/82.4**	**91.7/82.6**	**92.0**/82.7	**90.9/81.7**

The experimental results are shown in Table 13. It can be seen that in cross-modal scenarios, when the ego vehicle and the collaborator are equipped with sensors of different modalities and the correspondingly significant different encoders, UniMTO still demonstrates strong domain adaptation capabilities, achieving SOTA performance in one-to-one adaptation, many-to-one adaptation, and zero-shot adaptation. Furthermore, in Hetero Scenario 8, UniMTO is the only method with both AP0.5 and AP0.7 higher than No Fusion, further demonstrating the effectiveness of UniMTO.

D Simultaneous Collaboration From Mixed Unseen Domains

Table 14: Detection performance and computational overhead under simultaneous collaboration from mixed unseen domains. The fusion method is AttFuse.

Method	Params (M)	Peak memory (MB)	Latency (ms)	FLOPs (G)	AP0.5	AP0.7
NegoCollab[21]	16.8	4098.3	1075.4	462.9	58.8	49.5
PLDA[15]	34.5	4122.6	946.5	846.6	54.6	47.7
PnPDA[17]	8.1	4696.7	991.3	200.7	62.9	44.3
GenComm[16]	7.5	4086.4	979.7	92.4	55.8	47.7
STAMP[18]	6.9	4085.7	1017.5	128.7	56.8	48.8
MPDA[19]	18.6	4100.9	1543.0	427.5	60.2	49.8
PolyInter[20]	84.6	4185.1	956.1	262.3	59.5	50.1
UniMTO (Ours)	9.8	4113.9	1013.1	353.4	**66.1**	**54.0**

We further evaluate simultaneous collaboration on V2X-Real under Hetero Scenario 3. Three collaborators are active in the same scene and respectively use the unseen sd1, vn4, and pp8 encoders,

their featutres are jointly fused after adaptation. All results are measured with batch size 1 on an RTX 3090. Parameters and FLOPs count the adapter only, peak memory measures end-to-end inference, and latency includes preprocessing, forward, and postprocessing. As shown in Table 14, UniMTO remains effective when multiple collaborators from mixed unseen domains are simultaneously active, achieving the best performance among all compared methods while maintaining computational overhead comparable to existing methods. Communication efficiency is not the focus of this work, all methods transmit complete BEV features under the same protocol and therefore have identical communication volume.

E Leave-one-family-out Evaluation

We further conduct a leave-one-family-out experiment on V2V4Real. All methods are trained with pp4, pp8, sd1, and sd2, while the entire VoxelNet family (vn4 and vn8) is excluded from training and used only at inference. As shown in Table 15, UniMTO achieves the best performance on both unseen VoxelNet encoders among all compared methods. Notably, UniMTO is the only method that clearly outperforms No Fusion on both vn4 and vn8, whereas the other methods perform similarly to or worse than No Fusion. These results further demonstrate UniMTO's zero-shot generalization to an encoder family entirely excluded during training.

Table 15: Leave-one-family-out zero-shot adaptation results on the unseen VoxelNet family (AP0.5).

Method	vn4	vn8
No Fusion	56.6	56.6
NegoCollab[21]	53.0	56.8
PLDA[15]	55.2	56.1
PnPDA[17]	1.6	1.2
GenComm[16]	56.5	48.0
STAMP[18]	56.5	56.0
MPDA[19]	43.1	56.6
PolyInter[20]	56.0	56.5
UniMTO (Ours)	**61.0**	**64.8**

F Computational and Latency Overhead

In reality, due to the limited resources on the vehicle side, real-time performance is particularly important for collaborative perception. Therefore, we tested and compared the number of model parameters, inference time, and detection performance of all methods. As shown in Figure 7, since all other methods follow a one-to-one design, the number of deployed model parameters increases linearly with the number of new collaborators, which is undoubtedly disastrous and unacceptable in reality. In contrast, since UniMTO is a many-to-one domain adapter, only one is needed, which greatly reduces the deployment load. At the same time, as shown in Figure 8, compared with other methods, UniMTO not only achieves the best performance but is also competitive in inference time. Specifically, compared to PolyInter (inference time: 144.7ms, AP0.7: 43.3), UniMTO (inference time: 164.42ms, AP0.7: 46.9) has a 3.6% higher of AP0.7, while its inference time is only 14% longer.

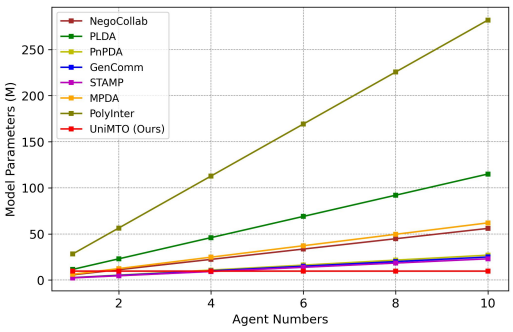


Figure 7: Comparison of model parameters when adding new heterogeneous collaborators.

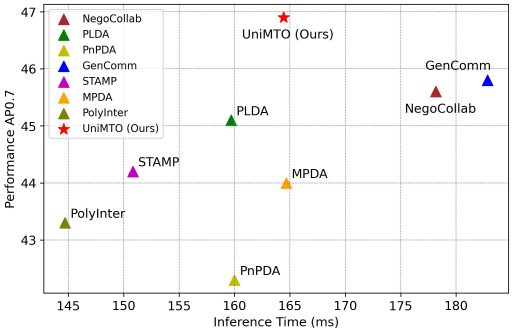


Figure 8: Efficiency-performance trade-off on V2V4Real. The fusion method is V2X-ViT.

G Limitations and Future Work

UniMTO is validated for 3D detection with task-compatible BEV features, aligned coordinates and compatible spatial coverage, and a fixed channel dimension; resolution differences can be interpolated,

while channel mismatches require projection and fine-tuning. Evaluation covers limited encoder families, sensor distributions, and agent scales, and CCCL may weaken under severe point-cloud distribution shifts. We do not claim universal zero-shot adaptation. Future work will explore channel-agnostic interfaces and architectures, sensors, tasks, and large-scale real-world collaborations.

H Detailed Values in Figure 4

Detailed values in Figure 4 are shown in Table 16, Table 17, and Table 18.

Table 16: Detailed values in Figure 4, the collaborator's encoder is sd1.

Method/Metric	AP0.5/AP0.7					
localization error std (m)	0.0	0.1	0.2	0.3	0.4	0.5
No Fusion	56.6/37.0					
NegoCollab[21]	69.7/43.5	68.9/42.0	66.5/39.4	63.2/37.7	60.2/36.6	57.9/36.1
PLDA[15]	56.6/37.1	57.2/37.3	56.9/37.1	56.7/36.8	56.3/36.8	56.0/36.5
PnPDA[17]	59.8/11.1	59.3/11.1	57.9/10.9	56.0/10.8	54.2/10.6	52.7/10.4
GenComm[16]	69.0/43.8	68.5/41.5	65.9/38.5	62.2/36.8	59.2/35.8	56.2/35.2
STAMP[18]	70.3/**45.0**	69.7/**42.9**	**67.0**/39.5	62.9/37.4	59.3/36.3	56.8/35.6
MPDA[19]	67.6/38.9	67.1/38.6	65.3/38.0	63.4/37.5	61.6/37.2	**60.0**/37.0
PolyInter[20]	67.6/42.4	67.1/40.9	65.0/38.2	62.0/36.5	59.4/35.6	57.3/34.8
UniMTO (Ours)	**70.4**/44.5	**69.7**/42.7	66.5/**40.5**	**64.2/38.3**	**62.3/37.5**	59.6/**37.0**

Table 17: Detailed values in Figure 4, the collaborator's encoder is vn8.

Method/Metric	AP0.5/AP0.7					
localization error std (m)	0.0	0.1	0.2	0.3	0.4	0.5
No Fusion	56.6/37.0					
NegoCollab[21]	69.7/44.0	69.0/42.0	66.5/39.5	63.3/37.9	60.3/36.8	58.3/36.3
PLDA[15]	58.5/39.2	58.4/38.8	58.1/38.3	57.8/38.0	57.3/37.7	57.0/37.5
PnPDA[17]	60.8/12.6	60.3/12.7	58.8/12.5	56.7/12.2	54.7/11.9	52.8/11.8
GenComm[16]	69.9/43.0	69.2/40.9	66.3/38.1	62.7/36.5	59.5/35.6	57.3/35.0
STAMP[18]	**71.7**/42.6	**70.5**/40.9	**67.5**/38.3	63.4/36.6	60.2/35.4	57.6/34.9
MPDA[19]	65.7/37.8	65.0/37.7	63.7/37.4	62.2/37.2	60.7/37.0	59.5/36.7
PolyInter[20]	67.6/41.8	67.1/39.5	64.6/36.4	61.4/34.4	58.6/33.6	56.2/33.1
UniMTO (Ours)	70.7/**44.6**	70.0/**42.0**	67.4/**40.5**	**63.9/39.1**	**61.2/37.9**	**60.2/37.2**

Table 18: Detailed values in Figure 4, the collaborator's encoder is pp4.

Method/Metric	AP0.5/AP0.7					
localization error std (m)	0.0	0.1	0.2	0.3	0.4	0.5
No Fusion	56.6/37.0					
NegoCollab[21]	70.2/43.0	69.4/41.4	66.7/38.9	63.5/37.6	60.7/36.6	58.4/36.0
PLDA[15]	56.6/37.1	56.3/36.8	56.0/36.6	55.5/36.5	55.3/36.2	55.2/36.0
PnPDA[17]	59.5/12.4	59.3/12.3	57.9/11.7	55.9/11.3	54.2/11.1	52.4/11.1
GenComm[16]	70.1/44.0	69.0/41.8	66.4/38.7	63.0/37.0	60.0/36.2	57.8/35.6
STAMP[18]	70.3/43.7	69.5/41.5	66.6/37.6	62.4/35.4	59.0/34.2	56.0/33.6
MPDA[19]	68.1/39.1	67.3/38.7	65.4/38.0	63.5/37.6	61.8/37.3	60.3/37.1
PolyInter[20]	67.7/41.5	67.3/40.0	65.3/37.3	62.0/35.5	59.1/34.5	56.7/34.0
UniMTO (Ours)	**70.7/44.5**	**69.8/42.0**	**67.9/39.7**	**65.0/38.6**	**62.8/37.6**	**60.4/37.1**

I More Visualizations

More visualizations of detection results on both V2V4Real and V2X-Real are shown in Figure 9 and Figure 10. It can be seen that all other methods suffer from varying degrees of missed detections and false detections, while only UniMTO can accurately detect all targets without false detections.

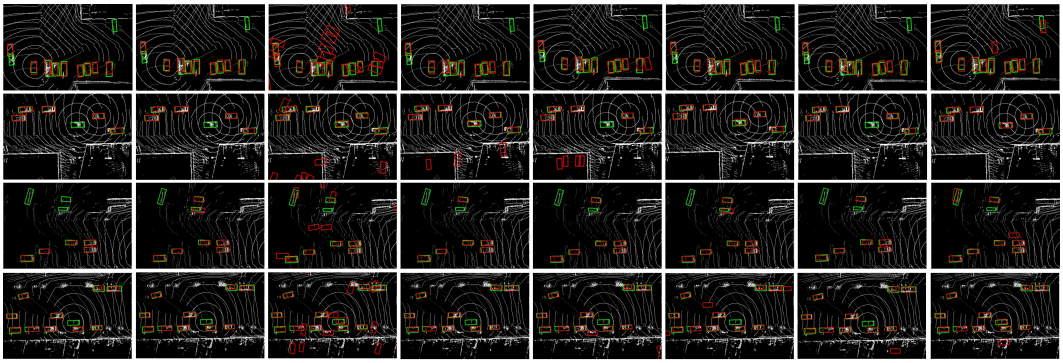


Figure 9: Visualization on V2V4Real. Green and red boxes denote ground truths and detection results.

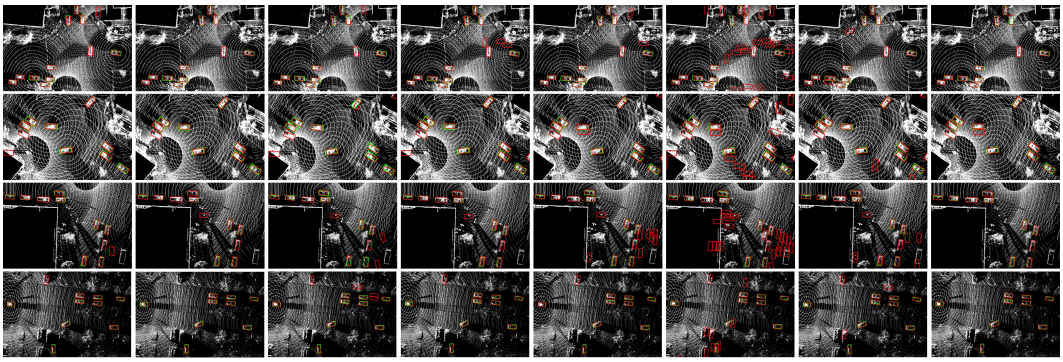


Figure 10: Visualization on V2X-Real. Green and red boxes denote ground truths and detection results.

More visualizations of features and its corresponding pattern and content are shown in Figure 11. As can be seen, due to the different encoders, the features and patterns of the ego vehicle and the collaborator differ significantly. Furthermore, the corresponding patterns and contents also exhibits clear distinguishability, proving the effectiveness of the separation concept in UniMTO. The visualization results further demonstrate the effectiveness and efficiency of UniMTO.

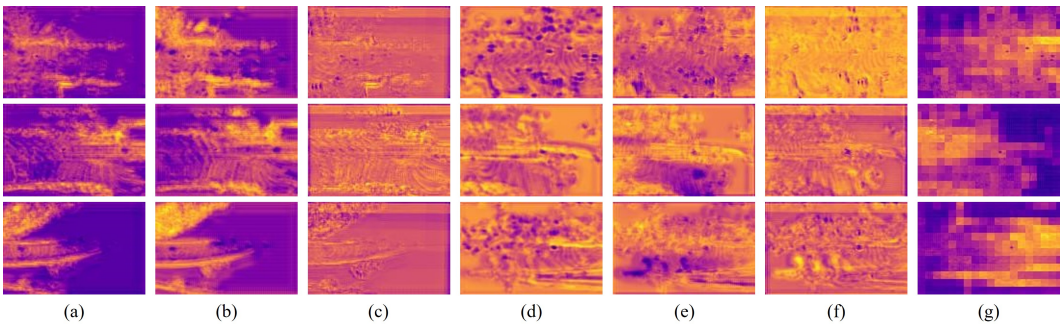


Figure 11: Visualization results in Hetero Scenario 2 on V2V4Real. The collaborator's encoder is sd1. (a) Ego's feature. (b) Ego's pattern. (c) Ego's content. (d) Collaborator's feature before domain adaptation. (e) Collaborator's pattern. (f) Collaborator's content. (g) Collaborator's feature after domain adaptation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and Section 1 accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix E discuss the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our code will be open-sourced to the public after the paper is officially published.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 specify all the training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it is too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information of computing resources in Section 4, and the execution time in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work conforms to the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Maybe there are some potential societal consequences of our work, none of which we feel must be specifically highlighted.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all external assets used in the paper (such as code, data, and models) are properly credited, with their licenses and terms of use clearly stated and strictly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release the code and related documentation upon the official publication of the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.