RSSA: Robust Semantic and Spatial Aligner for Collaborative Perception

Zhihao Yang[1] **Zhiyu Xiang**[1,2]* **Peng Xu**[1] **Tianyu Pu**[1]
Kai Wang[1] **Eryun Liu**[1] **Dongping Zhang**[3] **Yong Ding**[1]
[1]Zhejiang University
[2]Zhejiang Provincial Key Laboratory of Multi-Modal Communication Networks and Intelligent Information Processing
[3]China Jiliang University
yangzhihao@zju.edu.cn, xiangzy@zju.edu.cn

Abstract

V2X collaborative perception improves single-vehicle perception by aggregating multi-view features from collaborative agents. However, existing methods typically rely on simple spatial warping to align the collaborative features, which can cause semantic misalignment during fusion. Meanwhile, these deterministic warping-based methods are highly sensitive to the positional noises. In this paper, we propose a plug-in module called Robust Semantic and Spatial Aligner (RSSA), which performs semantic and probabilistic spatial transformations for robust feature alignment. Specifically, a Semantic Transformation (SemT) block modulates feature channels based on the observed relative pose to correct semantic inconsistencies. Subsequently, a Probabilistic Spatial Transformation (P-SpaT) block performs probabilistic warping by sampling from the Gaussian-modeled relative pose space. A Relative Transformation Augmentation (RT-Aug) strategy is further introduced to augment the relative poses and facilitate the training of the RSSA. Extensive experiments on both simulated and real-world datasets demonstrate that RSSA improves the state-of-the-art methods by a large margin under the configuration of both clean and noisy positions.

1 Introduction

Environmental perception is critical for autonomous driving. However, single-vehicle perception [11, 24, 21, 6, 13, 12, 2] is inherently limited by occlusion and restricted sensing range. V2X collaborative perception effectively addresses these limitations by aggregating information from multiple collaborative agents. These complementary multi-view features enable the ego vehicle to see through occlusions and extend its detection range, significantly enhancing driving safety.

Most collaborative perception methods follow a broadcast-based paradigm [14, 4, 18, 16]. In this paradigm, each collaborative

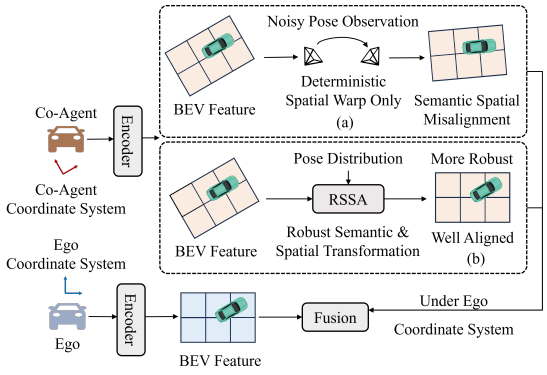

Figure 1: Comparison between the traditional spatial warping and our proposed RSSA for BEV features.

*Corresponding author.

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

agent extracts local BEV features, packages them with its pose, and then broadcasts them to the ego vehicle. Upon receiving the message, the ego vehicle spatially warps the collaborative features to its own coordinate system for subsequent feature fusion.

Although efficient, this paradigm suffers from two critical limitations. First, it overlooks the semantic misalignment caused by the spatial warping. In fact, deep features such as [6] inherently encode view-dependent geometric attributes. While spatial warping successfully aligns the collaborative features' position to the ego vehicle's, the semantic representations fail to adapt, which leads to suboptimal fusion results, as illustrated in Figure 1(a).

Second, the above deterministic feature alignment process is sensitive to the positional noises. Existing works attempt to mitigate the impact of the noisy position through downstream processing such as consensus matching [9, 23] or robust fusion modules [19], which increases the computing complexity. In fact, most modern positioning systems on-board provide location uncertainty information. Making full use of it can potentially improve the precision of feature alignment.

In this paper, we propose a plug-in module called Robust Semantic and Spatial Aligner (RSSA), which performs semantic and probabilistic spatial transformations for robust feature alignment, as illustrated in Figure 1(b). Specifically, RSSA incorporates a Semantic Transformation (SemT) block to correct semantic misalignment by modulating feature channels based on the observed relative pose. Subsequently, it employs a Probabilistic Spatial Transformation (P-SpaT) block to implement probabilistic warping, which samples relative poses based on the pose uncertainty. Furthermore, a Relative Transformation Augmentation (RT-Aug) strategy is introduced to improve the model's robustness against complex positional variations. We conduct extensive experiments on two representative V2X datasets: the simulated dataset V2XSet [19] and the real-world dataset V2X-Real [15]. Experimental results demonstrate that RSSA consistently boosts the performance of existing state-of-the-art methods under both clean and noisy position settings.

In summary, the main contributions are summarized as follows:

1. We propose RSSA, a plug-in module that performs semantic and probabilistic spatial transformations for robust collaborative feature alignment. It can effectively mitigate the semantic misalignment and reduce the sensitivity to positional noises for collaborative perception.
2. We introduce RT-Aug, an effective strategy that augments the relative poses with a random virtual coordinate system to boost the robustness of RSSA.
3. We conduct extensive experiments on both simulated and real-world datasets. Experimental results demonstrate that RSSA improves the state-of-the-art methods by a large margin.

2 Related Work

2.1 Paradigms in Collaborative Perception

As categorized in SparseAlign [22], collaborative perception methods can be mainly divided into two paradigms: handshake and broadcast. The handshake paradigm [19, 23] involves a two-round communication process. The ego vehicle first sends a request message with its pose to collaborative agents. These agents then transform their local raw data (e.g., point clouds) into the ego vehicle's coordinate system before feature extraction. While this early alignment ensures all features are learned in a unified coordinate frame, it suffers from high latency caused by its two-round communication, prohibitive computational overhead and high sensitivity to pose noises in feature fusion.

In contrast, the broadcasting paradigm [14, 4, 7, 18, 9, 16] is more efficient and widely adopted. In this scheme, each collaborative agent extracts features in its local coordinate system and broadcasts the features along with its pose. The ego vehicle then aligns the received features to its own coordinate system using a non-learnable spatial warping before fusion. While this approach is practical for its single round of communication, it suffers from semantic misalignment caused by spatial warping. That is to say, although the spatial positions of the features have been transformed, the view-dependent semantic information embedded within the features has not been effectively converted. A more realistic scenario is the existence of both semantic and spatial misalignment when positional noises

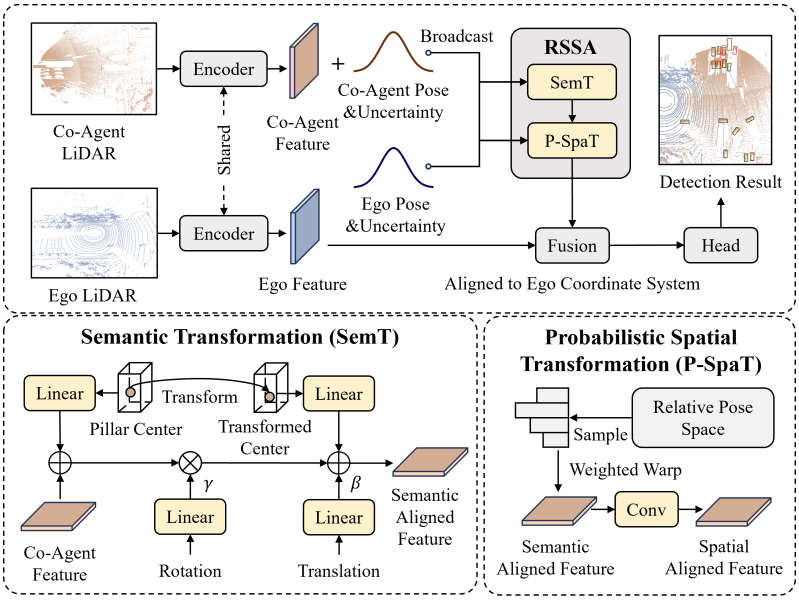

Figure 2: The overall collaborative perception pipeline and the architecture of our proposed RSSA.

are considered. In this paper, we focus on addressing these concerns by inventing semantic and probabilistic spatial transformations.

2.2 Feature Fusion in Collaborative Perception

A variety of intermediate fusion methods have been proposed to aggregate features from collaborative agents. AttFuse [20] introduces an agent-wise attention mechanism to adaptively weight features from different collaborators. DiscoNet [7] proposes a knowledge distillation framework where an early-fusion model acts as a teacher to improve the collaborative perception performance of the student network. Transformer-based architectures, such as V2X-ViT [19], have been adopted to effectively model the interactions between features from collaborative agents. CoBEVT [18] further improves upon this by introducing a fused axial attention to capture both local and global spatial relationships. To improve robustness against pose noise, CoAlign [9] designed a pose-graph optimization module before fusion. ERMVP [23] enhances communication efficiency through feature sampling and includes a spatial calibration module to handle localization errors. CoSDH [16] presented a hybrid fusion scheme to improve robustness against low communication bandwidth.

While effective, the performance of these fusion methods is critically limited by the quality of the upstream feature alignment. In this paper, we propose RSSA, a plug-in module designed for seamless integration with existing fusion networks to boost the collaborative detection accuracy.

3 Method

3.1 Overview

The overall collaborative perception pipeline and the architecture of RSSA are illustrated in Figure 2. Each agent extracts a Bird's-Eye-View (BEV) feature map from its local data and packages it with the pose and the uncertainty for broadcasting. Upon receiving the message, the ego vehicle employs the Robust Semantic and Spatial Aligner (RSSA) module to align the collaborative feature. The RSSA module consists of two sequential blocks. First, a Semantic Transformation (SemT) block corrects the semantic misalignment according to the observed poses. Subsequently, a Probabilistic Spatial Transformation (P-SpaT) block performs a probabilistic spatial warping guided by the pose uncertainty. The aligned feature is then fed into the fusion module to generate the fused feature. Finally, a detection head is used to obtain the perception results.

3.2 Semantic Transformation

The Semantic Transformation (SemT) block actually performs a semantic pre-alignment prior to the spatial transformation. The components of this block is illustrated in Figure 2.

Transformation Encoding. To incorporate the crucial spatial information, we need to encode the received relative poses. Considering the 6-DoF poses $p_i, p_j \in \mathbb{R}^6$ for the ego vehicle i and collaborative agent j, we first parameterize them into the corresponding pose matrices $P_i, P_j \in SE(3)$ and then calculate the relative transformation matrix $\Gamma_{j\to i}$:

$$\Gamma_{j\to i} = [R_{j\to i}, T_{j\to i}] = P_i \cdot P_j^{-1}, \tag{1}$$

where $R_{j\to i}$ and $T_{j\to i}$ represent the rotation matrix and translation vector. These components are then passed through linear layers to obtain the modulation vectors γ and β:

$$\gamma = \text{Linear}(\text{vec}(R_{j\to i})) \tag{2}$$

$$\beta = \text{Linear}(\text{norm}(T_{j\to i})), \tag{3}$$

where vec denotes flattening the rotation matrix and the norm function scales the translation vector based on the detection range.

Position Embedding. To provide spatial context, position embeddings are generated for both the collaborative agent j and ego vehicle i. Let C_j denotes the coordinates of the BEV pillar centers in the collaborative agent's frame, the source position embedding PE_j is produced by passing C_j through a linear layer:

$$PE_j = \text{Linear}(C_j). \tag{4}$$

Similarly, the target position embedding PE_i is generated by transforming C_j to the ego vehicle's coordinate system with $\Gamma_{j\to i}$ and then passing it through a linear layer:

$$PE_i = \text{Linear}(\Gamma_{j\to i} \cdot C_j). \tag{5}$$

Feature Modulation. The feature modulation integrates the position embeddings and the transformation encodings to semantically pre-align the collaborative agent's feature F_j to the ego vehicle's coordinate system. First, the source position embedding PE_j is added to F_j for integrating initial spatial context:

$$F'_j = F_j + PE_j. \tag{6}$$

The modulation vectors γ and β are then used to modulate F'_j:

$$F^{mod}_{j\to i} = \gamma \cdot F'_j + \beta. \tag{7}$$

Finally, the target position embedding PE_i is added to obtain the semantically aligned feature:

$$F^{sem}_{j\to i} = F^{mod}_{j\to i} + PE_i, \tag{8}$$

Through this transformation-aware feature modulation, the SemT block effectively pre-align the semantic information for the subsequent spatial shift of the feature.

3.3 Probabilistic Spatial Transformation

The co-agent's features are required to transform into the ego's frame before fusion. Existing methods simply regard the relative pose as deterministic quantities and apply it in a deterministic space transformation, which results in high sensitivity to positional noise. In this paper we explore the value of position uncertainty which can be readily obtained from the common localization systems, and encode this information into spatial transformation. We model the pose as Gaussian distribution and perform probabilistic warping by sampling from the relative pose space, as shown in Figure 2.

Relative Pose Space Construction. Sampling directly from the high dimensional joint 6DoF pose space is inefficient. To reduce the complexity, we leverage the structural decomposition of the relative transformation matrix $\Gamma_{j\to i}$:

$$\Gamma_{j\to i} = \begin{bmatrix} R_i^T R_j & R_i^T (t_j - t_i) \\ 0 & 1 \end{bmatrix}, \tag{9}$$

where R_i, t_i and R_j, t_j denote the rotation and translations for the ego vehicle i and collaborative agent j, respectively. Since $\Gamma_{j\to i}$ is determined by: the orientations of both agents and their relative position difference, we construct a compact relative pose space $\mathcal{S}$ as:

$$\mathcal{S} = \{r_i, r_j, \Delta t\}, \text{ with } \begin{aligned} r_i &\sim \mathcal{N}(r_i^{obs}, \sigma_{r,i}^2), \\ r_j &\sim \mathcal{N}(r_j^{obs}, \sigma_{r,j}^2), \\ \Delta t &\sim \mathcal{N}(\Delta t^{obs}, \sigma_{t,i}^2 + \sigma_{t,j}^2). \end{aligned} \tag{10}$$

where $\mathcal{N}(,)$ is Gaussian distribution, r_i^{obs} and r_j^{obs} denote the observed rotation components of the poses, Δt^{obs} denotes the observed relative translation, and $\sigma_{r,i}, \sigma_{r,j}$ and $\sigma_{t,i}, \sigma_{t,j}$ represent the corresponding rotation and translation uncertainties for the ego vehicle i and collaborative agent j, respectively.

Probabilistic Warping. To approximate the distribution efficiently, we employ Quasi-Monte Carlo [1] sampling to generate K pose samples from the relative pose space $\mathcal{S}$, which include $K-1$ samples and the mean observation itself.

The spatially-aligned feature $F_{j\to i}^{spa}$ is then obtained by weighted aggregation of the features warped by these pose samples, followed by a convolutional layer:

$$F_{j\to i}^{spa} = \text{Conv}\left(\sum_{k=1}^{K} w_k \cdot \text{Warp}(F_{j\to i}^{sem}, \Gamma_{j\to i}^{k})\right), \tag{11}$$

$$w_k = \frac{\exp(p_k)}{\sum_{m=1}^{K}\exp(p_m)}, \tag{12}$$

where $\text{Warp}(\cdot)$ denotes the spatial warping, and p_k is the probability of the k-th sample. By leveraging this uncertainty-aware probabilistic spatial warping, the P-SpaT block effectively mitigates the interference of pose noise and significantly enhances the robustness of feature alignment.

3.4 Relative Transformation Augmentation

In typical V2X scenarios, structured layouts result in limited relative spatial relationship, which inherently restricts the generalization ability of RSSA. To address this geometric bias, we propose the Relative Transformation Augmentation (RT-Aug) strategy, as illustrated in Figure 3. It introduces a random virtual coordinate system v to diversify the relative poses.

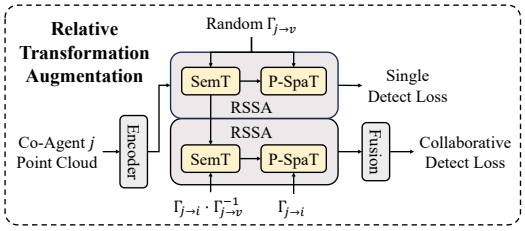

Figure 3: Illustration of the Relative Transformation Augmentation (RT-Aug) strategy.

Specifically, RT-Aug involves a two-stage process. First, a random transformation $\Gamma_{j\to v}$ is sampled to transform the collaborative feature F_j to a virtual coordinate system v:

$$F_{j\to v}^{sem} = \text{SemT}(F_j, \Gamma_{j\to v}), \tag{13}$$

$$F_{j\to v}^{spa} = \text{P-SpaT}(F_{j\to v}^{sem}, \Gamma_{j\to v}). \tag{14}$$

$F_{j\to v}^{spa}$ is supervised by the single-agent detection loss L_{single} in the virtual coordinate system to ensure the correctness of the virtual transformation.

Second, the intermediate feature $F_{j\to v}^{sem}$ is aligned to the ego vehicle's frame for subsequent feature fusion. SemT applies a compensation transformation to shift semantics from the virtual coordinate system to the ego-vehicle's, while P-SpaT applies the original relative transformation to warp the feature from the collaborative agent's coordinate system to the ego-vehicle's:

$$F_{j\to i}^{sem} = \text{SemT}(F_{j\to v}^{sem}, \Gamma_{j\to i}\cdot\Gamma_{j\to v}^{-1}), \tag{15}$$

$$F_{j\to i}^{spa} = \text{P-SpaT}(F_{j\to i}^{sem}, \Gamma_{j\to i}). \tag{16}$$

Finally, $F_{j\to i}^{spa}$ is fused with the ego feature and supervised by the collaborative detection loss L_{coll}.

We jointly train the RSSA and the collaborative perception backbone in an end-to-end manner. The total training objective is defined as the sum of the aforementioned losses:

$$L_{total} = L_{single} + L_{coll}, \tag{17}$$

where both L_{single} and L_{coll} follow the standard detection loss, employing Focal Loss [8] for classification and Smooth L1 loss for regression.

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate our method on the public V2XSet [19] and V2X-Real [15] datasets. V2XSet is a large-scale simulated collaborative dataset involving vehicles and infrastructure, with a detection range of $x \in [-140.8m, 140.8m], y \in [-38.4m, 38.4m], z \in [-5m, 3m]$. V2X-Real is a large-scale real-world collaborative dataset featuring multiple vehicles and infrastructures, with a detection range of $x \in [-100.8m, 100.8m], y \in [-38.4m, 38.4m], z \in [-5m, 3m]$.

Implementation Details. Our framework is implemented in PyTorch [10] and trained on two NVIDIA RTX 3090 GPUs. We use the Adam[5] optimizer with an initial learning rate of 0.001, which is decayed by a factor of 0.1 at epochs 10 and 20. The model is trained for a total of 40 epochs, and early stopping is used to select the best epoch. We employ PointPillars [6] as the 3D backbone to extract BEV features, with a pillar size of 0.4m x 0.4m. For the P-SpaT module, the number of pose samples K is set to 9.

We conduct experiments under two settings: Clean Setting and Noisy Setting. In the Clean Setting, models are trained and evaluated without adding artificial pose noise. In the Noisy Setting, we add Gaussian noises to the agents' poses to simulate various levels of pose errors. Specifically, during training, the translation noise is sampled from $\mathcal{N}(0, \sigma_t^2)$ and applied to x and y coordinates, with σ_t uniformly sampled from the range $[0.2m, 0.4m]$. For rotation, noise is sampled from $\mathcal{N}(0, \sigma_r^2)$ and applied to the yaw angle, with σ_r uniformly sampled from $[0.2°, 0.4°]$. During inference, we evaluate the models with varying noise levels, with the standard deviations for translation and rotation noise range from 0 to 1.0 m/°.

For the RT-Aug strategy, the virtual transformation is constructed by randomly sampling a yaw angle uniformly from $[-180°, 180°]$ and a z-axis translation uniformly from $[-0.5m, 0.5m]$. During inference, the virtual transformation is simply set to an identity transformation.

Evaluation Metrics. We use both BEV and 3D Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7 to evaluate the collaborative 3D object detection performance.

4.2 Quantitative Evaluation

Performance Comparison under Clean Setting. We evaluate the effectiveness of RSSA by incorporating it into several state-of-the-art methods. The results under the clean settings are shown in Table 1, where the broadcast and broadcast+RSSA paradigms are the main focus of our evaluation, while the handshake is for reference.

We can observe that the handshake paradigm consistently outperforms the broadcast in this experiment. For instance, on the V2XSet dataset with CoBEVT, handshake surpasses broadcast by a large margin of 16.8% in BEV AP@0.7. The key difference is that handshake performs alignment on raw point clouds, ensuring features' generation in a unified coordinate system with the ego's, thereby eliminating the semantic inconsistencies. This gap becomes even more pronounced when 3D AP is considered, where significant height differences between the infrastructure and the vehicle sensors exist. This comparison highlights the importance of feature semantic alignment, which is fundamentally overlooked by existing broadcast-based methods.

Our RSSA is designed to bridge this gap. By incorporating RSSA, the performance of the broadcast methods are substantially improved. Specifically, RSSA, with its lightweight design (only 1.3M parameters), boosts the BEV and 3D AP@0.7 of CoBEVT by 14.18% and 16.09% on V2XSet, and by 5.78% and 7.41% on V2X-Real, respectively. These results demonstrate that RSSA effectively corrects the semantic misalignment on the feature-level and can achieve comparable performance

Table 1: Performance comparison with state-of-the-art methods on the V2XSet and V2X-Real datasets under the clean setting. The best results are highlighted in **bold**. H and B denote handshake and broadcast, respectively. The values in red indicate the performance improvement of our RSSA.

Method		V2XSet		V2X-Real	
		BEV AP@0.5/0.7	3D AP@0.5/0.7	BEV AP@0.5/0.7	3D AP@0.5/0.7
Att fuse [20]	H	93.20/85.46	92.07/75.73	74.29/58.83	70.35/34.92
	B	84.47/68.42	81.13/55.09	68.79/53.24	61.80/26.04
	B+RSSA	(+7.00/+15.19)	(+8.58/+16.17)	(+5.14/+7.12)	(+4.34/+9.49)
		91.47/83.61	**89.71/71.26**	**73.93/60.36**	**66.14/35.53**
CoBEVT [18]	H	93.07/85.85	91.98/76.11	76.43/61.10	72.31/37.39
	B	86.80/69.05	82.89/55.62	72.85/55.71	65.78/31.00
	B+RSSA	(+4.79/+14.18)	(+7.17/+16.09)	(+2.97/+5.78)	(+2.37/+7.41)
		91.59/83.23	**90.06/71.71**	**75.82/61.49**	**68.15/38.41**
ERMVP [23]	H	93.72/86.65	92.81/76.63	75.96/59.69	71.72/37.08
	B	86.61/67.89	82.29/53.41	71.38/53.80	64.29/29.90
	B+RSSA	(+5.00/+15.65)	(+7.73/+18.96)	(+5.43/+8.92)	(+5.34/+8.25)
		91.61/83.54	**90.02/72.37**	**76.81/62.72**	**69.63/38.15**
CoSDH [16]	H	95.15/89.61	93.78/80.87	78.68/65.93	75.03/40.97
	B	92.18/85.08	86.93/70.60	77.99/63.53	61.72/26.49
	B+RSSA	(+1.31/+3.39)	(+3.28/+5.97)	(+3.54/+4.99)	(+12.56/+14.38)
		93.49/88.47	**90.21/76.57**	**81.53/68.52**	**74.28/40.87**

to that of the handshake paradigm. For more detailed analysis of RSSA's computational cost and latency, please refer to Appendix C.5.

Notably, on the real-world V2X-Real dataset, RSSA-enhanced methods can even outperform the handshake paradigm. For instance, ERMVP+RSSA surpasses the handshake configuration by 3.03% in BEV AP@0.7 and 1.07% in 3D AP@0.7. This is because real-world datasets inherently contain localization and calibration noise, even in the clean setting. The handshake paradigm is highly sensitive to such noise when aligning raw point clouds. In contrast, RSSA adaptively learns a robust transformation to handle these real-world spatial inconsistencies.

Performance Comparison under Noisy Setting. We further evaluate the robustness of our method against pose errors by introducing varying levels of positional noises, as shown in Figure 4. As expected, the performance of the handshake paradigm drops dramatically as noise level increases. For example, on the V2XSet with CoBEVT, the BEV AP@0.7 of handshake paradigm decreases rapidly from 81.2% to 42.6%. In contrast, the broadcast baseline methods exhibits a slower degradation and performs better during most of the time, as it relies less on precise spatial alignment given the pre-existing semantic misalignment.

Our RSSA further outperforms the broadcast baselines across all noise levels. In low-noise settings, the significant performance gain validates the effectiveness of SemT in correcting semantic misalignment. As noise increases, RSSA maintains superior robustness thanks to the probabilistic warping of P-SpaT. Specifically, at the noise level of 1.0 m/° on V2X-Real, CoSDH+RSSA increases the BEV AP@0.7 by 9.8% compared to the baseline. These results confirm the effectiveness of RSSA in addressing both semantic and spatial misalignment.

4.3 Ablation Study

We conduct comprehensive ablation studies on the V2XSet dataset under noisy settings, with CoBEVT as the baseline to validate the effectiveness of each proposed components.

Effectiveness of each component. As shown in Figure 5, adding only the SemT block to the broadcast baseline results in a substantial performance boost of +4.9% in BEV AP@0.5 and +2.4% in 3D AP@0.5 at the zero noise level. This confirms that correcting the semantic misalignment is crucial for effective collaborative perception. However, as the SemT-only baseline still relies on a

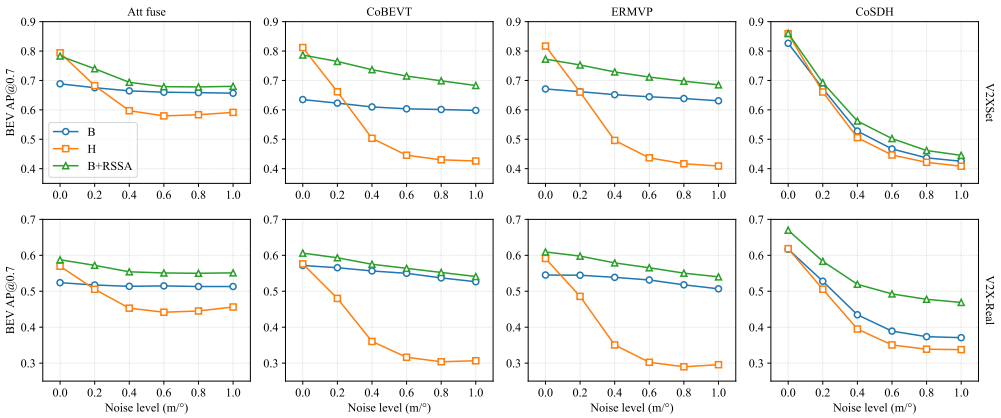

Figure 4: Performance comparison with state-of-the-art methods on the V2XSet (top row) and V2X-Real (bottom row) datasets under various noise levels. Columns from left to right are: Att fuse [20], CoBEVT [18], ERMVP [23], CoSDH [16], respectively. B and H denote Broadcast and Handshake, respectively.

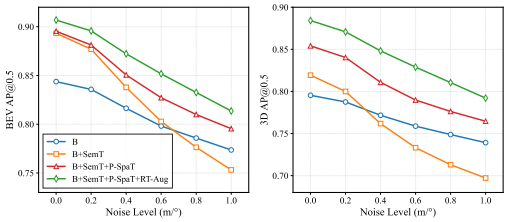

Figure 5: Ablation study of RSSA components.

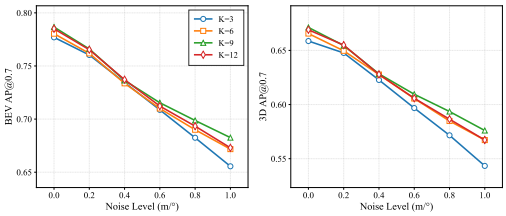

Figure 6: Ablation study of sampling number.

deterministic spatial warping, its performance degrades sharply as noise increases, similar to the handshake paradigm.

By incorporating the P-SpaT block, the performance curve becomes much flatter along the noise axis. At a high noise level of 1.0 m/°, P-SpaT brings a substantial performance gain of +6.8% in 3D AP@0.5 over the SemT-only baseline.

Finally, the inclusion of RT-Aug (the full RSSA model) yields a consistent performance improvement across all noise levels. This gain is attributed to the enhanced generalization of the SemT module. The RT-Aug strategy diversifies the training data, which guides the model to handle more complex geometric relationships, such as cases of turning maneuvers or on uneven ground.

Sampling number in P-SpaT. We investigate the impact of the number of pose samples K in our P-SpaT module. As shown in Figure 6, increasing the number of samples K from 3 to 6 and 9 yields noticeable performance gains. Further increasing the K to 12 leads to a slight performance drop. Therefore, we choose $K = 9$ as it offers the best trade-off between accuracy and efficiency.

4.4 Generalization Study

We conduct generalization studies on the V2XSet dataset under noisy settings with CoBEVT as the baseline.

Generalization to inaccurate uncertainty. We investigate the scenario where the provided uncertainty from the localization system is not accurate. This is simulated by multiplying the provided uncertainty with a scale coefficient $\tau \in [0.5, 1.5]$. As shown in Figure 7, the performance slightly decreases as the uncertainty becomes inaccurate and consistently outperforms the broadcast baseline.

Generalization to unseen noise type. To assess RSSA's generalization capability to different types of pose noise, we evaluate its performance with Laplace noise during inference. As shown

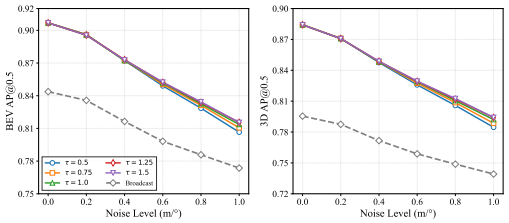

Figure 7: Generalization study of uncertainty.

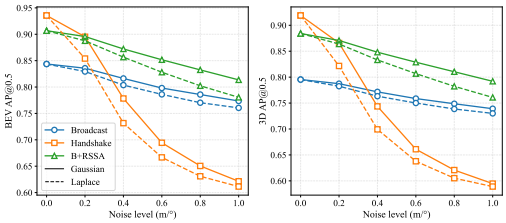

Figure 8: Generalization study of noise type.

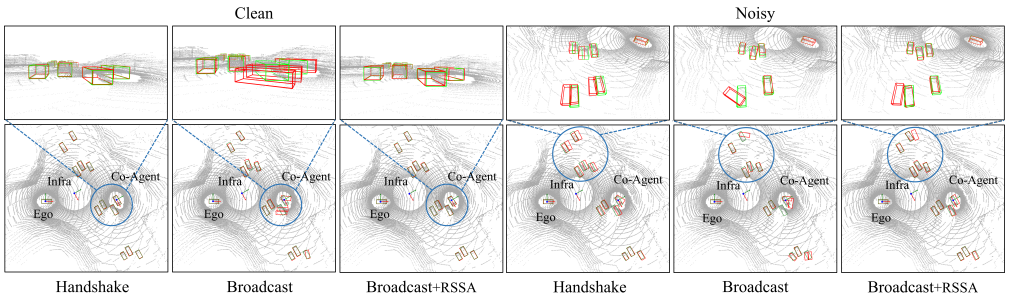

Figure 9: Visualization of detection results under the clean (left three columns) and noisy (right three columns) settings. For each setting, the upper row shows the zoomed-in view of the region highlighted by the blue circle in the lower BEV view. Green and red boxes represent ground truth and predictions, respectively.

in Figure 8, while all methods exhibit a performance drop when fed with the unseen Laplace noise, RSSA consistently maintains the highest performance across all noise levels.

4.5 Qualitative Evaluation

We conduct qualitative analysis on the V2XSet dataset using CoBEVT as the baseline.

Visualization of the detection results under clean setting. As shown in left three columns of Figure 9, the broadcast paradigm exhibits significant errors in orientation and height due to semantic misalignment. Specifically, for the region highlighted by the blue circle, the predicted object orientations are inconsistent with the ground truth. Similarly, in the 3D view, the object center heights are incorrectly estimated. In contrast, our RSSA correctly aligns both the orientation in the BEV plane and the height in 3D space, producing better collaborative 3D object detection results.

Visualization of the detection results under noisy setting. The right three columns of Figure 9 illustrate the results under the noise level of 1.0 m/°. The predictions from handshake paradigm suffers from significant spatial displacements and false detections due to raw point misalignment. In contrast, our RSSA demonstrates superior robustness by effectively mitigating spatial noise and maintaining correct semantics, yielding accurate detection results that closely match the ground truth.

5 Conclusion

In this paper, we proposed RSSA, a novel plug-in module to strengthen the modern broadcast-based collaborative perception methods. The module is designed to address the semantic and spatial misalignment problems inherent in the deterministic spatial warping employed by most collaborative methods. RSSA integrates two key components: the Semantic Transformation (SemT) block, which corrects semantic misalignment using the observed relative pose, and the Probabilistic Spatial Transformation (P-SpaT) block, which performs robust spatial warping via probabilistic sampling guided by the pose uncertainty. The RT-Aug strategy is further proposed to augment the relative poses of the training samples. Extensive experiments on both simulated and real-world datasets demonstrate that RSSA consistently improves the performance of the state-of-the-art methods by a large margin under clean and noisy conditions.

Acknowledgments

This work was supported by the Key Project of Natural Science Foundation of Zhejiang Province under Grant LZ26F010003, the Key Research & Development Plan of Zhejiang Province under Grants 2024C01010 and 2024C01017, the Joint R&D Program of the Yangtze River Delta Community of Sci-Tech Innovation under Grant 2024CSJGG01000, and the National Key Laboratory of Collective Intelligence & Collaboration (Open Fund Project No. QXZ25017101).

References

[1] Russel E Caflisch. Monte carlo and quasi-monte carlo methods. *Acta numerica*, 7:1–49, 1998.

[2] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1201–1209, 2021.

[3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[4] 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.

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

[6] 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.

[7] 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.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[9] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4812–4818. IEEE, 2023.

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[11] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[12] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020.

[13] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[14] 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.

[15] 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.

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

[17] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1155–1162. IEEE, 2021.

[18] 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.

[19] 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.

[20] 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.

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

[22] Yunshuang Yuan, Yan Xia, Daniel Cremers, and Monika Sester. Sparsealign: A fully sparse framework for cooperative object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22296–22305, 2025.

[23] 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.

[24] 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.

A Analysis of Semantic Misalignment

A.1 Mathematical Analysis

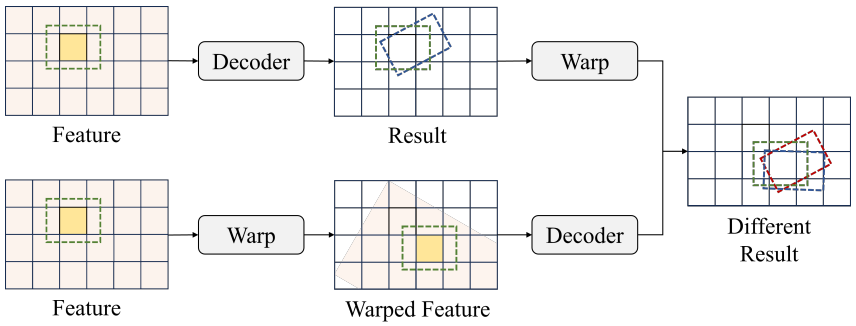

Figure 10: Comparison of detection results between different align methods. The yellow square represents the foreground feature, and the green box indicates the anchor. The top row (result-level alignment) yields the target result (blue box), whereas the bottom row (feature-level alignment) results in a misalignment (red box). The discrepancy confirms that spatial warping of BEV features causes semantic misalignment and leads to detection error, highlighting the importance and necessity of semantic alignment during coordinate transformation.

The motivation of RSSA is that spatial warping of BEV features in the broadcast paradigm causes semantic misalignment. To demonstrate this, we compare the detection results of two align methods: *result-level alignment* (correct alignment) and *feature-level alignment* (spatial warping in the broadcast paradigm), as illustrated in Figure 10.

Let $f(p)$ be the foreground local BEV feature at grid p and the detection head interprets $f(p)$ using an anchor with orientation $\mathbf{o}_{anc}$. The decoded center c and orientation o are:

$$c = p + \Delta p(f(p)), \quad o = o_{anc} \cdot \Delta o(f(p)). \tag{18}$$

Here, $\Delta p(f(p))$ represents the predicted spatial offset from the anchor center p to the object center, and $\Delta o(f(p))$ denotes the predicted orientation offset relative to the anchor's base orientation $\mathbf{o}_{anc}$.

For the *result-level alignment*, we apply a spatial transformation $T = (R, t)$ to the detection results:

$$c^* = R(p + \Delta p(f(p))) + t = (Rp + t) + R\Delta p(f(p)), \quad o^* = R \cdot (o_{anc} \cdot \Delta o(f(p))). \tag{19}$$

For the *feature-level alignment*, we apply spatial warping $\mathcal{W}$ to the feature map f before decoding. Since $\mathcal{W}$ only transform the grid location $p \to p' = Rp + t$ while keeping the local feature semantic $f(p)$ unchanged, the decoded results from the warped feature $\mathcal{W}(f; T)$ at p' are:

$$\hat{c} = p' + \Delta p(f(p)) = (Rp + t) + \Delta p(f(p)), \quad \hat{o} = o_{anc} \cdot \Delta o(f(p)). \tag{20}$$

By comparing $(\hat{c}, \hat{o})$ with the (c^*, o^*), we identify the problems of feature-level alignment:

1. Center error: Although spatial warping successfully maps the grid position p to $p' = Rp + t$, the predicted spatial offset $\Delta p(f(p))$ remains anchored to the original coordinate system. Consequently, the decoded center $\hat{c}$ fails to incorporate the rotation R into the predicted offset, resulting in a discrepancy of $c^* - \hat{c} = (R - I)\Delta p(f(p))$.
2. Orientation error: Similarly, while the grid position is spatially transformed to p', the semantic representation $f(p)$ remains invariant, result in a discrepancy of $o^* \cdot \hat{o}^{-1} = R$.

This discrepancy proves that spatial warping of BEV feature causes semantic misalignment. When $R \neq I$, this misalignment leads to substantial error in predicted object center and orientation, which demonstrate the importance and necessity of semantic alignment in collaborative perception.

A.2 Visual Analysis

Figure 11 provides a zoomed-in visualization of the Broadcast and Broadcast+RSSA paradigms under the clean setting, corresponding to a detailed portion of Figure 9 in the main paper.

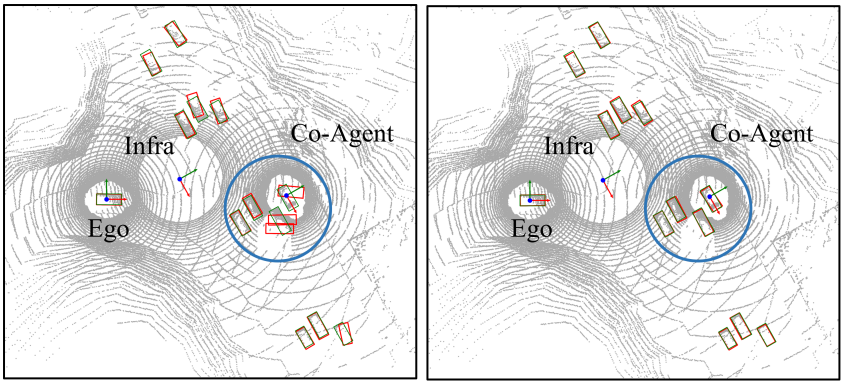

Figure 11: Visualization of detection results under the clean setting (zoomed-in portion of Figure 9). The incorrect detection result within the blue circle confirms that spatial warping causes semantic misalignment, leading to orientation and center errors. By resolving the semantic misalignment, RSSA yields more accurate detection results.

As shown in Figure 11, the detection results of broadcast paradigm within the blue circle exhibit significant error. These objects are originally parallel to the x-axis in the coordinate systems of the co-agent and infrastructure. However, according to our proof in Eq.(20), spatial warping only transforms the feature position but fails to align the feature semantic. Consequently, during fusion, the ego vehicle interprets the orientation semantic as being relative to its own coordinate system, resulting in incorrect predictions that are parallel to its own x-axis.

Furthermore, the predicted centers also exhibit a noticeable shift from the ground truth. This is consistent with the proved center error $c^* - \hat{c} = (R - I)\Delta p(f(p))$, confirming that the prediction centers are shifted due to the lack of rotational compensation for the spatial offset. In contrast, by aligning both semantics and spatial coordinates, our RSSA effectively corrects these errors and produces more accurate detection results, as shown on the right.

A.3 Direct Validation of SemT

To directly validate the ability of SemT, we analyze the detection results obtained from the features before SemT and after SemT using Broadcast+RSSA with CoBEVT [18] on V2XSet [19] under the clean setting. We match detection results with ground-truth boxes using an IoU threshold of 0.5 and report the mean center and orientation error over the matched results (True Positives), as shown in Table 2.

Table 2: Mean error before and after SemT on V2XSet under the clean setting.

Feature	Center error (m)	Orientation error (°)
Before SemT	0.52	45.93
After SemT	0.17	10.78

Even among TPs, SemT reduces the mean center error by 0.35 m and the mean orientation error by 35.15°. Before SemT, the detection results exhibit a large center error and a more pronounced orientation error, consistent with our mathematical analysis in Appendix A.1: spatial warping transforms feature locations but leaves the center-offset and orientation semantics in the source coordinate system. After SemT, both errors are substantially reduced, providing direct evidence that SemT learns the intended semantic transformation.

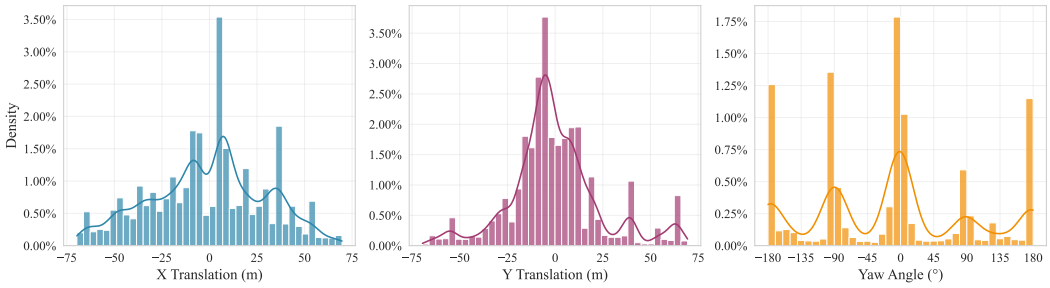

Figure 12: Statistical analysis of relative pose distributions on the V2XSet dataset. The yaw angle distribution (right) exhibits significant data bias with peaks at $0°, \pm 90°,$ and $\pm 180°$, highlighting the structured nature of urban scenes and the necessity of the RT-Aug strategy.

B Analysis of Relative Pose Distributions

As discussed in the section 3.4, V2X scenarios are often highly structured, which results in limited relative spatial relationships between agents. To empirically demonstrate this, we provide a statistical analysis of the relative poses between collaborative agents and the ego vehicle on the V2XSet [19] dataset, as illustrated in Figure 12.

The distributions of relative X and Y translations (left and middle) exhibit a wide range of values due to the diverse movement of agents. More importantly, the yaw angle distribution (right) also shows a varied spread resulting from maneuvers such as turns. As mathematically proven in Appendix A.1, whenever a relative rotation exists ($R \neq I$), spatial warping alone fails to align feature semantics, which empirically justifies the necessity of RSSA for semantic alignment.

Simultaneously, the yaw distribution reveals a distinct data bias, with pronounced peaks at $0°$, $\pm 90°$, and $\pm 180°$. This phenomenon is expected in structured urban environments, where agents typically move in parallel, opposite, or perpendicular directions at intersections. Such biased distribution can cause the network to overfit to these specific orientations, potentially hindering its ability to handle arbitrary relative angles, such as those encountered during complex turning maneuvers. To address this limitation, we proposed the RT-Aug strategy. By introducing a random virtual coordinate system during training, RT-Aug effectively diversifies the relative transformations, enabling the model to learn generalized feature alignment.

C Additional Experiments

C.1 Ablation Studies

Position Embeddings in SemT. To assess the contribution of the position embeddings in SemT, we compare RSSA with and without them on V2XSet [19] using CoBEVT [18] under the clean setting.

Table 3: Ablation of position embeddings on V2XSet with CoBEVT under the clean setting.

Method	BEV AP@0.7	3D AP@0.7
Broadcast+RSSA	83.23	71.71
Broadcast+RSSA, w/o position embeddings	81.69	67.87

As shown in Table 3, removing the position embeddings within SemT decreases 3D AP@0.7 by 3.84 points. This confirms that the position embeddings provide complementary global spatial context for semantic transformation rather than duplicating the relative-transformation encoding.

Isolated Impact of P-SpaT. To isolate the contribution of P-SpaT, we compare RSSA with and without P-SpaT on V2XSet using CoBEVT under the noisy setting.

Table 4: Ablation of P-SpaT on V2XSet with CoBEVT (BEV AP@0.7). Noise levels are given in m/°.

Method	0.2	0.4	0.6	0.8	1.0
RSSA w/o P-SpaT	72.74	70.52	68.26	66.33	64.41
RSSA w/ P-SpaT	76.46	73.66	71.50	69.88	68.28

As shown in Table 4, incorporating P-SpaT consistently improves performance across all noise levels. At the 1.0 m/° noise level, P-SpaT improves AP by 3.87 points, demonstrating the effectiveness of probabilistic spatial alignment.

Component-Level Ablation with ERMVP. To verify the effectiveness of our design, we conduct the same component-level ablation on V2XSet using ERMVP [23] under the noisy setting.

Table 5: Component-level ablation on V2XSet with ERMVP (BEV AP@0.5). Noise levels are given in m/°.

Method	0.0	0.2	0.4	0.6	0.8	1.0
Broadcast	85.84	85.24	83.48	81.71	79.94	78.20
Broadcast+SemT	89.09	87.48	82.48	77.75	74.46	71.75
Broadcast+SemT+P-SpaT	89.65	88.06	85.03	82.61	81.05	79.90
Broadcast+RSSA	89.84	88.73	86.49	84.68	82.92	81.33

As shown in Table 5, the component-level ablation trends with ERMVP are consistent with those observed using CoBEVT. At zero noise, SemT improves AP by 3.25 points, validating its contribution to semantic alignment. As noise increases, SemT alone degrades due to deterministic spatial warping, whereas P-SpaT substantially improves robustness, outperforming SemT by 8.15 points at the 1.0 m/° noise level. Finally, incorporating RT-Aug achieves the best performance across all noise levels, exceeding SemT+P-SpaT by 1.43 points at 1.0 m/°. These results confirm the effectiveness and generalizability of the RSSA components.

C.2 Robustness to Other Noise

Miscalibrated Uncertainty. We evaluate CoBEVT on V2XSet under zero-mean pose noise with the actual standard deviation fixed at 0.4 m/°. To simulate miscalibrated uncertainty, we provide P-SpaT with input uncertainty levels of 1.0 m/°, 2.0 m/°, and 4.0 m/°.

Table 6: Robustness to miscalibrated pose uncertainty (BEV AP@0.7).

Method	1.0 m/°	2.0 m/°	4.0 m/°
Handshake	50.32	50.32	50.32
Broadcast	61.00	61.00	61.00
Broadcast+RSSA	73.93	73.60	73.25

As shown in Table 6, although RSSA's performance gradually decreases as the input uncertainty becomes less accurate, it consistently outperforms both Broadcast and Handshake. Since Broadcast and Handshake do not use the input uncertainty, their performance remains unchanged across the three settings. Even with an input uncertainty of 4.0 m/°, RSSA remains 12.25 and 22.93 AP points higher than Broadcast and Handshake, respectively, demonstrating its robustness to miscalibrated uncertainty.

Biased Localization Errors. We further evaluate robustness to systematic localization bias on V2XSet with CoBEVT. We introduce non-zero-mean pose noise with the standard deviation fixed at 0.4 m/°, while setting the mean bias to 0.2 m/°, 0.4 m/°, and 0.6 m/°.

As shown in Table 7, Broadcast+RSSA achieves the best performance across all bias levels. At the largest mean bias of 0.6 m/°, it outperforms Broadcast and Handshake by 10.72 and 29.03 AP points, respectively, demonstrating its robustness to biased localization errors.

Table 7: Robustness to biased localization errors (BEV AP@0.7).

Method	0.2 m/°	0.4 m/°	0.6 m/°
Handshake	48.86	45.26	41.75
Broadcast	60.88	60.43	60.06
Broadcast+RSSA	73.13	72.24	70.78

Missing Data. We evaluate robustness to missing data on V2XSet with CoBEVT by randomly dropping collaborative agent's data with probabilities of 0.25, 0.50, and 0.75.

Table 8: Robustness to missing data (BEV AP@0.7).

Method	0.25	0.50	0.75
Handshake	77.81	74.03	69.19
Broadcast	63.78	64.00	64.30
Broadcast+RSSA	76.02	72.83	68.97

As shown in Table 8, Broadcast+RSSA achieves performance comparable to Handshake and consistently outperforms Broadcast across all missing probabilities. At the highest missing probability of 0.75, it remains comparable to Handshake while outperforming Broadcast by 4.67 AP points, demonstrating its robustness to missing data.

Temporal Desynchronization. We evaluate robustness to temporal desynchronization on V2XSet with CoBEVT by introducing time delays to collaborative agent's data from 100 ms to 400 ms.

Table 9: Robustness to temporal desynchronization (BEV AP@0.7).

Method	100 ms	200 ms	300 ms	400 ms
Handshake	68.52	56.34	53.04	52.43
Broadcast	61.45	59.57	59.18	58.88
Broadcast+RSSA	74.74	69.25	63.95	60.35

As shown in Table 9, the data desynchronization caused by time delays degrades collaborative perception performance. Nevertheless, Broadcast+RSSA achieves the best performance across all delay levels. At the largest delay of 400 ms, it outperforms Broadcast and Handshake by 1.47 and 7.92 AP points, respectively, demonstrating its robustness to temporal desynchronization.

C.3 Effect of Noisy Training

As commonly adopted in previous works, CoAlign [9] and V2X-ViT [19], we introduce synthetic pose noise during training for all evaluated methods. Clean training optimizes the model for accurate poses, whereas noisy training exposes it to pose errors. Consequently, the two training settings can produce different results even at zero additional test noise.

To quantify this trade-off, we evaluate CoBEVT+RSSA trained under the clean and noisy settings across different test noise levels, as shown in Table 10.

On V2XSet, noisy training sacrifices 4.58 AP points at zero noise but improves AP by 34.12 points at the 1.0 m/° noise level. On V2X-Real, it improves AP by 12.82 points at 1.0 m/°, with only a 0.88-point decrease at zero additional noise. Thus, noisy training provides substantial gains under high pose noise, with a marginal decrease at zero additional noise on V2X-Real.

Nevertheless, we agree that synthetic noise cannot fully represent real-world pose errors. On the other side, current carefully calibrated and constructed benchmark datasets often contain less pose noise in their clean settings than real-world scenarios. Therefore, noisy training remains valuable for practical deployment.

Table 10: Robustness evaluation of CoBEVT+RSSA under different training settings (BEV AP@0.7). Test noise levels are given in m/°.

Dataset	Training	0.0	0.2	0.4	0.6	0.8	1.0
V2XSet [19]	Clean	83.23	62.75	41.13	33.97	32.68	34.16
V2XSet	Noisy	78.65	76.46	73.66	71.50	69.88	68.28
V2X-Real [15]	Clean	61.49	51.75	41.34	38.71	39.28	41.28
V2X-Real	Noisy	60.61	59.32	57.51	56.38	55.23	54.10

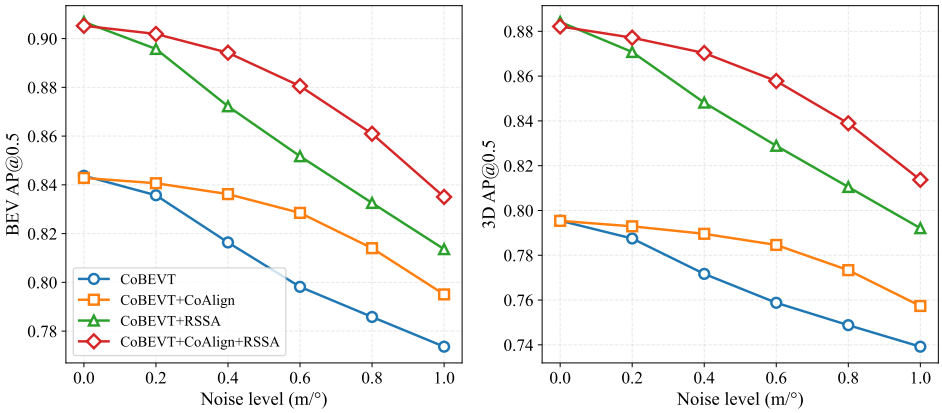

Figure 13: Performance comparison of CoAlign and RSSA using CoBEVT as the fusion module on the V2XSet dataset.

C.4 Complementarity with CoAlign

To demonstrate the universality of RSSA, we evaluate its complementarity with other robustness method, specifically the pose-graph optimization in CoAlign. As illustrated in Figure 13, both CoAlign and RSSA independently provide performance gains over the baseline, and their joint integration yields the best results across all noise levels. This improvement stems from the fact that CoAlign optimizes alignment at the pose level, while RSSA ensures robust and semantically consistent alignment at the feature level. These results prove that RSSA is a general plug-in module that can effectively enhance other robustness methods.

C.5 Model Complexity and Efficiency

Table 11: Comparison of model complexity, efficiency, and performance using CoBEVT as fusion module on the V2XSet dataset.

Method	Parameters	Flops	Latency	BEV AP@0.7	
	(M)	(G)	(ms)	Clean	Noisy (1.0 m/°)
Broadcast	10.51	251.05	97.18	69.05	59.82
Broadcast+RSSA	11.84	281.27	111.76	**83.23**	**68.28**
RSSA	1.33	30.22	14.58	+14.18	+8.46

We analyze the complexity and efficiency of RSSA to demonstrate its practicality as a lightweight plug-in. Table 11 compares the parameters, GFlops, latency, and detection performance with and without RSSA using CoBEVT as fusion module on the V2XSet dataset.

As shown in Table 11, RSSA is highly efficient. Since the module consists primarily of simple linear layers and convolutional layers, the parameter increment is only 1.33 M. Furthermore, thanks to the parallel execution of the probabilistic warping, RSSA introduces a marginal computational overhead of 30.22 GFlops and only 14.58 ms of additional latency. With this minimal cost, our module effectively achieves semantic and robust spatial alignment, obtaining substantial performance

gains of +14.18% in the clean setting and +8.46% under high pose noise level (1.0 m/°) on BEV AP@0.7. These results confirm that RSSA provides a superior trade-off between detection accuracy and computational efficiency.

D Qualitative Evaluation on V2X-Real Dataset

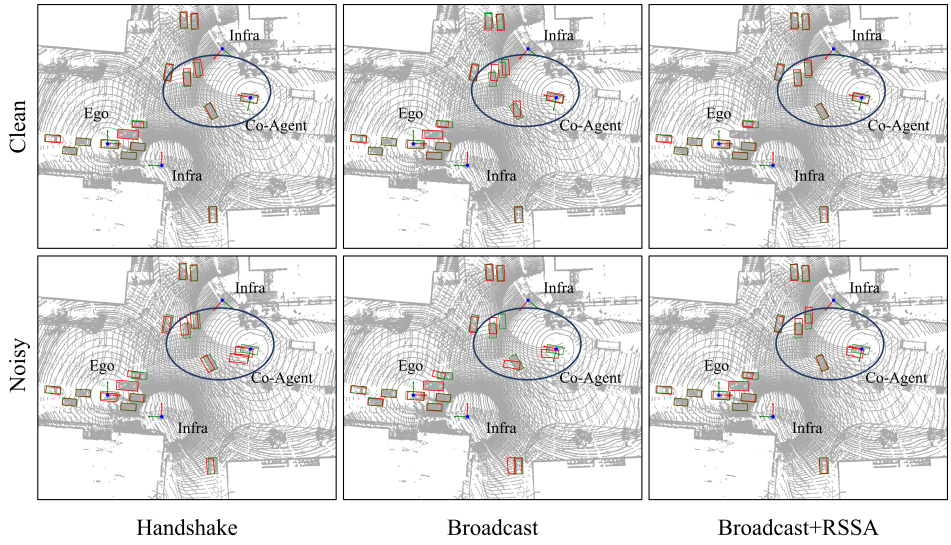

Figure 14: Visualization of detection results on the V2X-Real dataset using CoBEVT as fusion module across clean (top) and noisy (bottom) settings. The blue circles highlight that the broadcast paradigm suffers from semantic misalignment caused by spatial warping, and the handshake paradigm is sensitive to noise, resulting in more false positives. Our RSSA provides the most robust and accurate detection by performing semantic transformation and probabilistic spatial transformation.

We provide additional visualization results on the real-world V2X-Real dataset using CoBEVT as baseline. As shown in Figure 14, the observations are highly consistent with those on the simulated V2XSet dataset. In the clean setting, the broadcast paradigm exhibits significant inaccuracies within the blue circle due to semantic misalignment, while RSSA yields precise results by effectively correcting this misalignment.

In the noisy setting, the broadcast paradigm remains inaccurate due to persistent semantic errors. Meanwhile, the handshake paradigm produces more false positives because it is extremely sensitive to pose noise during raw data alignment. In contrast, our RSSA maintains superior robustness and delivers the most accurate detection results, further confirming its effectiveness and reliability in complex real-world scenarios.

E Limitations

While our proposed RSSA demonstrates significant improvements in collaborative perception, there are a few limitations. Firstly, our validation is solely conducted on the 3D object detection. While RSSA's core contribution lies in robust feature alignment, its applicability and performance on other perception tasks, such as semantic segmentation or tracking, remain to be explored. Secondly, RSSA fundamentally relies on position information provided by the localization system and is not designed to operate under unknown pose conditions. Future work will aim to extend RSSA to a wider range of perception tasks and investigate its performance and adaptability in scenarios with completely unknown relative poses.

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: The main claims in the abstract and introduction accurately reflect the contributions and scope of the paper, providing a concise and faithful overview of the work presented.

   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: We have discussed the limitations of the work in the appendix.

   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: This paper does not include theoretical result.

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 would be 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 device in section 4 and time of execution in appendix C.5.

   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: We 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 discussed.

    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: We have cited the creators or original owners of assets used in the paper and mentioned and respected the license and terms of use.

    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 open code and related documentation after the paper is officially published.

    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 and formatting purposes in the paper.

    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.