CHARM: Collaborative Harmonization Across Arbitrary Modalities for Modality-Agnostic Semantic Segmentation

Lekang Wen[1], Jing Xiao[2*], Liang Liao[3], Jiajun Chen[1], Mi Wang[1],

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University
[2]School of Artificial Intelligence, Wuhan University
[3]Hangzhou Institute of Technology, Xidian University
wenlk3@whu.edu.com, jing@whu.edu.com, liaoliang1@xidian.edu.cn, jiajunchen@whu.edu.com, wangmi@whu.edu.com

Abstract

Modality-agnostic Semantic Segmentation (MaSS) aims to achieve robust scene understanding across arbitrary combinations of input modality. Existing methods typically rely on explicit feature alignment to achieve modal homogenization, which dilutes the distinctive strengths of each modality and destroys their inherent complementarity. To achieve cooperative harmonization rather than homogenization, we propose CHARM, a novel complementary learning framework designed to implicitly align content while preserving modality-specific advantages through two components: (1) Mutual Perception Unit (MPU), enabling implicit alignment through window-based cross-modal interaction, where modalities serve as both queries and contexts for each other to discover modality-interactive correspondences; (2) A dual-path optimization strategy that decouples training into Collaborative Learning Strategy (CoL) for complementary fusion learning and Individual Enhancement Strategy (InE) for protected modality-specific optimization. Experiments across multiple datasets and backbones indicate that CHARM consistently outperform the baselines, with significant increment on the fragile modalities. This work shifts the focus from model homogenization to harmonization, enabling cross-modal complementarity for true harmony in diversity.

1 Introduction

Multimodal Semantic Segmentation (MSS) aims to leverage complementary features from multiple visual modalities to achieve robust scene understanding (Valada, Mohan, and Burgard 2020; Gandhi et al. 2023; Hegde et al. 2025), particularly under adverse conditions, *e.g.*, heavy fog, nighttime, and rain. For instance, in autonomous driving scenarios under rainy conditions, RGB-based perception often suffers from substantial degradation due to reduced visibility (Liao et al. 2022, 2023; Wang et al. 2023). In contrast, auxiliary modalities such as LiDAR and Depth sensors can provide geometry-aware cues that compensate for the perception limitations of RGB, thereby improving driving safety. Most existing MSS methods follow an *X*+*A* paradigm, where *X* denotes as a designated primary modality and *A* represents one (Tan et al. 2024; Dong et al. 2024; Wei et al. 2025; Zhang et al. 2025; Yang et al. 2025) or multiple auxiliary modalities (Zhang et al. 2023a; Brödermann et al. 2023; Li et al. 2025; Wan et al. 2024; Reza, Prater-Bennette, and Asif 2024). However, this paradigm inherently prioritizes the primary modality, which suppresses the contribution of auxiliary modalities, leading to significant performance degradation when the primary modality is unavailable.

Modality-agnostic Semantic Segmentation (MaSS) has emerged to enable robust performance under diverse and incomplete modality configurations. Early efforts tackled multimodal imbalance by employing modality dropout (Zhang et al. 2023b, 2024; Shi et al. 2024; Maheshwari, Liu, and Kira 2024), which randomly omits input sources during training. However, it may be insufficient to ensure effective representation learning for non-dominant or fragile modalities. Recent methods (Wang et al. 2022a; Zheng, Lyu, and Wang 2024; Zheng et al. 2024a,b, 2025) have shifted toward feature alignment, enforcing cross-modal feature consistency through explicit constraints, *e.g.*, KL divergence. MAGIC (Zheng et al. 2024a) groups modalities into robust and fragile sets and aligns their features with a cosine-based loss. Any2Seg (Zheng, Lyu, and Wang 2024) incorporates semantic guidance from vision-language models to enhance alignment, while AnySeg (Zheng et al. 2025) distills all modality features toward informative representation, improving performance in fragile-modal combinations. Although these alignment-based methods have shown to better handle combinations involving robust modalities, their reliance on explicit alignment often suppress modality-specific characteristics, leading to homogenized representations and reduced cross-modal complementarity.

As illustrated in Fig. 1(a), different modalities could capture distinct aspects of the same scene, resulting in inherent misalignment. For example, the building missed in LiDAR while the Lane marking is only available in RGB. Explicit alignment-based methods as shown in (b) homogenizes these multimodal features by enforcing convergence across inherently divergent distributions. This process weakens modality-specific strengths, as RGB's rich information is diluted to match achromatic patterns of LiDAR for the building case, and LiDAR's precise spatial cues are suppressed in the car case as shown in (d). This limitation motivates us to pursue a novel approach that preserves the distinct characteristics of each modality while enabling ef-

*Corresponding author

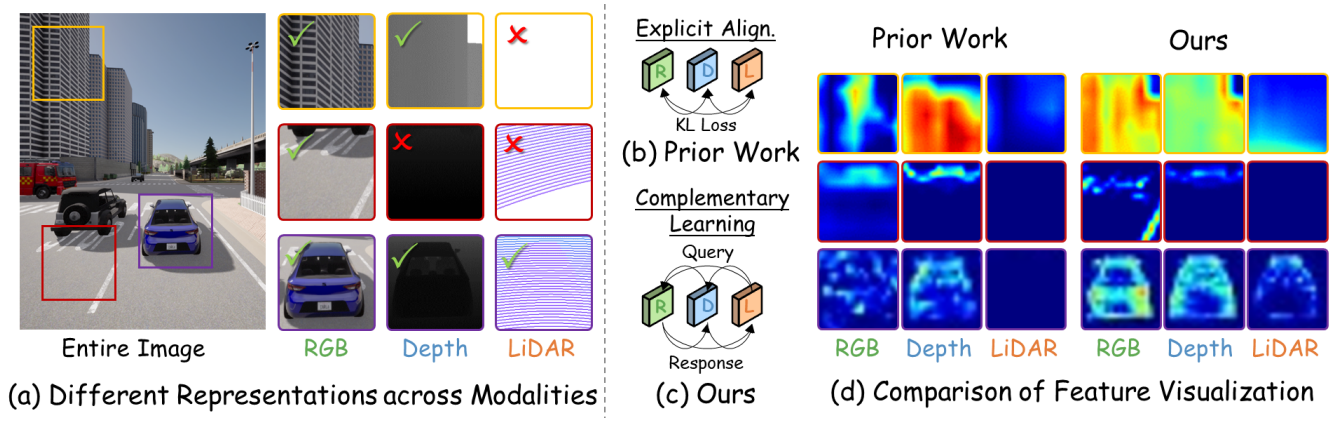

Figure 1: Example of modality misalignment and comparison of designed methods for MaSS. (a) Examples of content variations in different modalities; (b) Prior works use explicit alignment (*e.g.*, KL divergence) that forces feature homogenization; (c) Our approach employs complementary learning through query and response to discover modality-interactive correspondences; (d) features of RGB and LiDAR are suppressed in prior work while in our work the features show their boosted specific information.

fective collaboration, namely to harmonize modalities under two principles: 1) modalities should coordinate through mutual understanding of each other, enabling the system to jointly model comprehensive scene information across modalities; 2) the complementary features of each modality should be actively stimulated and reinforced during the multi-modal interaction, rather than suppressed by uniform alignment constraints.

In this work, we propose **C**ollaborative **H**armonization across **AR**bitrary **M**odalities (CHARM), a MaSS framework that enables synergistic feature interaction while preserving modality-specific strengths. The core of CHARM lies in two aspects according to the above principles. Firstly, a fundamental unit, named modality Mutual Perception Unit (MPU), is proposed to explore the complementarity among modalities through cross-modal attention, where each modality simultaneously serve as queries and contexts to the others, enabling the discovery of modality-interactive correspondences at each feature scale. Second, two pathways are designed to systematically balance collaborative learning of all modalities (principle 1) with individual enhancement (principle 2): Collaboration Learning strategy (CoL) for joint optimization that dynamically adapts cooperation based on each modality's robustness, and Individual Enhancement strategy (InE) that provides protected learning spaces with the robustness guidance for individual modalities to stimulate their full potential.

The main contributions of this work are threefold:

- We identify the fundamental limitation of modality homogenization in existing explicit alignment strategy for MaSS on suppressing the modal-specific features, and propose a cooperative paradigm that enables effective complementarity across modalities;
- We designed a complementary learning framework, *i.e.*, CHARM, with MPU to explore the complementary across modalities, and two strategies of CoL for collaborative learning of all modalities and InE for individual enhancement by mutually learning from other modality;
- Extensive experiments demonstrate that CHARM significantly improves the effectiveness of each modality and their various combinations, ensuring robust MaSS performance that consistently surpasses advanced baselines.

2 Methodology

2.1 Problem Formulation of MaSS

Following the common setting, we consider a modality set $\mathcal{M}$ comprising M modalities. Let $\mathcal{I} = \{I_m\}_{m\in\mathcal{M}}$ represent the full multimodal input, and S denote the corresponding segmentation labels. During training, all modalities are assumed to be accessible, while during inference, only a subset $\mathcal{I}^{\text{sub}} \subseteq \mathcal{I}$ is available due to sensor failures, environmental variability, or cost constraints. The goal of MaSS is to learn a segmentation model that maintains robust and accurate under arbitrary modality combinations, particularly in the presence of missing modalities.

Typically, models trained with full multimodal supervision suffers significant performance degradation when certain modalities are missing. To mitigate this issue, existing approaches employ a twofold solution: (1) random modality dropout during training to simulate diverse incomplete modality scenarios, generating input subsets $\mathcal{I}_{\text{sub}}$, and (2) cross-modal feature alignment to ensure representation consistency across modalities. The model Φ is then optimized to handle various modal combinations by minimizing:

$$\min_{\Phi} \mathcal{L}_{\text{MaSS}} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}(\Phi(\mathcal{I}_{\text{sub}}), S) + \lambda_{\text{align}} \sum_{(m,m')\in\mathcal{P}} \mathcal{L}_{\text{align}}(f_m, f_{m'})), \quad (1)$$

where $\mathcal{P} = \{(m, m')|m, m' \in \mathcal{M}, m \neq m'\}$ represents all distinct modality pairs, f_m and $f_{m'}$ denote the features from modalities m and m', $\mathcal{L}_{\text{seg}}$ is the cross-entropy loss for segmentation, $\mathcal{L}_{\text{align}}$ is the similarity loss for multimodal feature alignment, and λ_{seg}, λ_{align} are balancing coefficients.

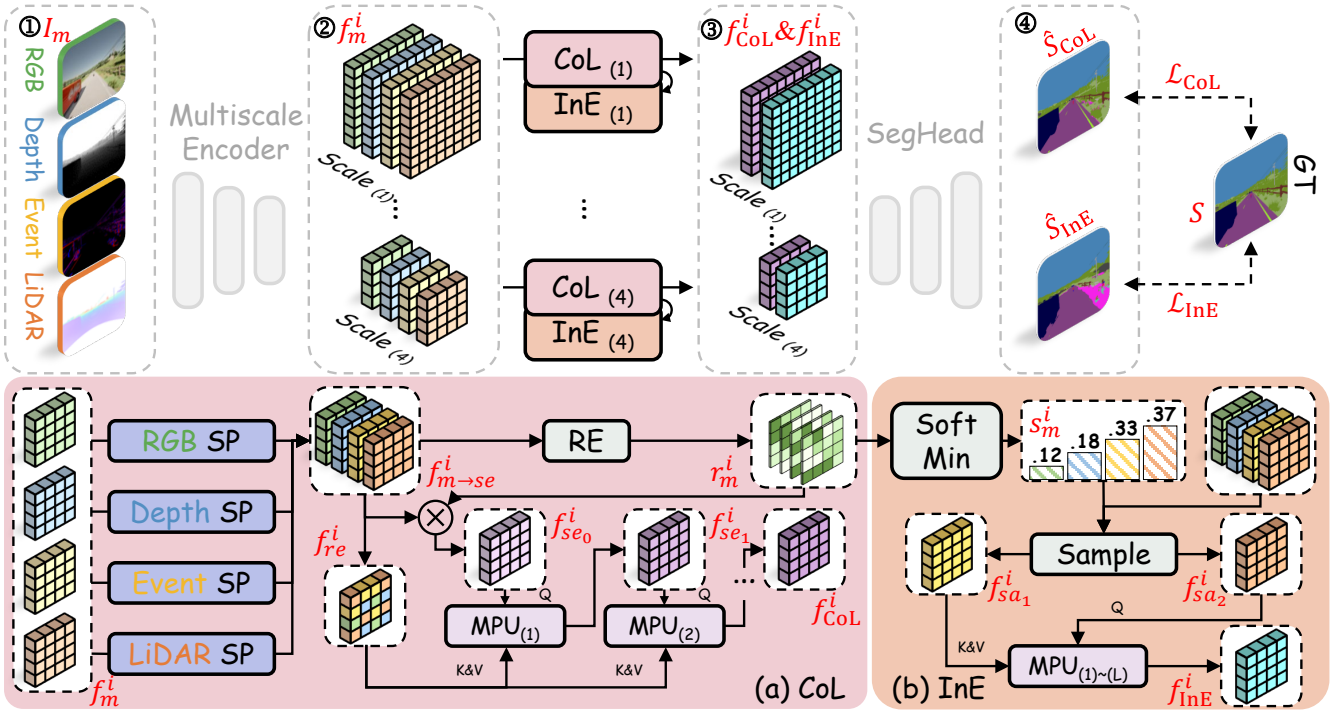

Figure 2: The overall framework of CHARM. At each scale, the encoder extracts multimodal features. CoL and InE are coupled when training and added between scales to fuse them progressively and input into the SegHead for segmentation.

2.2 Framework Overview

The framework of the proposed CHARM is shown in Fig. 2, where we formulate MaSS as a cooperative harmonization problem and solve it with a two-pathway synergistic process. MPU is the core module that can accept arbitrary numbers and types of multimodal features as queries and contexts, enabling each querying modality to perceive the content of others during learning. It facilitates the discovery of modality-interactive correspondences, resulting in implicit alignment without explicit constraints. These aligned modalities can, in turn, complement and enhance each other.

To leverage MPU effectively, CHARM integrates two cooperative pathways: CoL uses extracted multimodal features as queries and contexts to mine modality-interactive correspondences between multiple modalities, enabling model to learn complementary information and maximize synergy across modalities; InE employs individual modal features as queries and contexts to mine interactive responses within single modalities. Although CoL effectively enhances complementary capabilities, the issue of modal imbalance potentially suppress the fragile modalities. In contrast, InE provides a protective learning space for all modalities, thereby improving the individual potential of each modality.

Taking four modalities of RGB (R), Depth (D), Event (E), and LiDAR (L) modalities as an example, CHARM packs all input images I_m into a mini-batch for efficient parallel computation. A shared-weight encoder F extract features for each modality independently across each scale:

$$\{f_m^i\}_{i=1}^4 = F(I_m). \tag{2}$$

It is important to note that CoL and InE operate in couples: the feature f_m^i from each modality is processed through both pathways to generate a modal-collaborative feature f_{CoL}^i and a modal-enhanced feature f_{InE}^i. There features are then fed into a segmentation head to produce the modal-collaborative prediction $\hat{S}_{\text{CoL}}$ and the modal-enhanced prediction $\hat{S}_{\text{InE}}$ for joint optimization, respectively.

2.3 Mutual Perception Unit

To discover modality-interactive correspondences, MPU enables mutual perception across modalities through iterative query, where each modality's features serve as both queries (Q) and contexts (K, V) for the others. Designed to be modality-agnostic, MPU can robustly handle arbitrary modality combinations during inference. For multi-modal imagery, a windowed attention is designed to balance a large cross-modal receptive field with computational efficiency.

To further enhance cross-modal interaction, multiple MPU blocks are employed in CHARM, where each block follows a specific structure shown in Fig. 3. Specifically, $f_{m\to se}^i$ is partitioned into J windowed modal semantic features $\{x_m^{i,j}\}_{j=1}^J$, while $f_{se_{l-1}}^i$ is partitioned into J windowed semantic features $\{x_{se}^{i,j}\}_{j=1}^J$. Each $\{x_m^{i,j}\}_{j=1}^J$ has modality-specific weight matrices for generating Key and Value, thus the generation process of Q, K, V is formulated as:

$$\begin{aligned} Q_{i,j} &= x_{se}^{i,j} W^Q, \\ K_{i,j} &= \{x_m^{i,j} W_m^K\}_{m\in\mathcal{M}}, \\ V_{i,j} &= \{x_m^{i,j} W_m^V\}_{m\in\mathcal{M}}. \end{aligned} \tag{3}$$

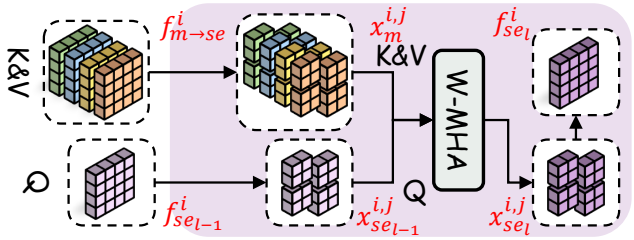

Figure 3: Structure of Mutual Perception Unit (MPU).

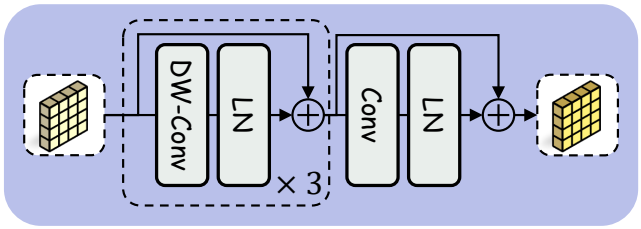

Figure 4: Structure of Semantic Projector (SP).

These serve as inputs for the Window Multi-Head Attention (W-MHA) (Liu et al. 2021), outputting interacted windowed semantic features $\{x_{se}^{i,j}\}$, which are finally assembled through reverse partition to obtain the semantic feature $f_{se_l}^i$. Shifted window partitioning is introduced for cross-window connections by shifting during adjacent MPU blocks. The separable Q, K, V relationships prevent any single modality from dominating the learning process, ensuring balanced multimodal interaction. L MPU blocks are embedded within both pathways described in the following at each scale. In CoL, MPU optimally integrates simultaneous multi-modal inputs, while in InE, it enables each modality to learn modality-interactive correspondences.

2.4 Collaboration Learning Strategy

As shown in Fig. 2(a), in CoL, the modal features $\{f_m^i\}_{m\in\mathcal{M}}$ at each scale are first projected to a semantic space via Semantic Projectors (SP), yielding $\{f_{m\to se}^i\}_{m\in\mathcal{M}}$. Each SP is a modality-specific sequential block consisting of three depth-wise convolutions with kernel sizes of 11×11, 7×7, and 3×3, followed by a 1×1 point-wise convolution, as described in Fig. 4. To quantify the reliabilit of each modality at every pixel, a Robustness Evaluator (RE), implemented as a 1×1 point-wise convolution, is used to compute the pixel-wise robustness scores $\{r_m^i\}_{m\in\mathcal{M}}$. These scores are then used to compute the initial semantic feature $f_{se_0}^i$ via a robustness-weighted summation:

$$f_{se_0}^i = \sum_{m\in\mathcal{M}} r_m^i f_{m\to se}^i, \tag{4}$$

which serves as the Query for the first MPU block. To reduce computational cost and modal dependency, modality-reassembled feature f_{re}^i is randomly selected from the corresponding position across all modality-specific semantic features, serving as the Key and Value for MPU block to enable fine-grained, cross-model interaction. Through iterative processing across MPU blocks, the semantic features are progressively refined, ultimately generating the fused modal feature f_{CoL}^i. This design encourages complementary information exchange by enforcing cross-modal interactions, allowing modalities to compensate for each other's deficiencies in a content-aware and position-sensitive manner.

2.5 Individual Enhancement Strategy

As shown in Fig. 2(b), in InE, the modality robustness scores $\{r_m^i\}_{m\in\mathcal{M}}$ are averaged to obtain scalar values $\{\bar{r}_m^i\}_{m\in\mathcal{M}}$, which are then normalized through *SoftMin* to derive sampling probabilities $\{s_m^i\}_{m\in\mathcal{M}}$:

$$s_m^i = \textit{SoftMin}(\bar{r}_m^i),\ m\in\mathcal{M}. \tag{5}$$

Subsequently, fragile modality-biased sampling is employed based on $\{s_m^i\}_{m\in\mathcal{M}}$ to extract complete modal semantic features $f_{sa_1}^i$ and $f_{sa_1}^i$ that serve as Q, K, V for MPU blocks, generating the enhanced modal feature f_{InE}^i. This fragile modality-biased learning strategy emphasizes the potential of fragile modalities by providing them with more learning opportunities. Since the input modal features are complete and do not incorporate information from other modalities, InE implements protective learning for individual modalities. The combination of individual modal features and mutual perception facilitates modalities to discover complementary information from each other and promotes implicit alignment across modalities.

2.6 Model Training and Inference

As output features from CoL and InE have the same shape, they are packaged as input to SegHead for parallel segmentation. The loss functions $\mathcal{L}_{\text{CoL}}$ and $\mathcal{L}_{\text{InE}}$ both use cross-entropy loss for supervised training:

$$\begin{aligned}\mathcal{L}_{\text{CoL}} &= -\sum_{p\in S} S(p)\log\left(\hat{S}_{\text{CoL}}(p)\right),\\ \mathcal{L}_{\text{InE}} &= -\sum_{p\in S} S(p)\log\left(\hat{S}_{\text{InE}}(p)\right),\end{aligned} \tag{6}$$

where p is the pixel index for segmentation map S. The MaSS optimization is then simplified as:

$$\min_{\Phi}\mathcal{L}_{\text{MaSS}} = \lambda_{\text{CoL}}\mathcal{L}_{\text{CoL}} + \lambda_{\text{InE}}\mathcal{L}_{\text{InE}}, \tag{7}$$

where $\mathcal{L}_{\text{CoL}}$ and $\mathcal{L}_{\text{InE}}$ denote the collaborative learning loss and modal enhancement loss, respectively. λ_{CoL} and λ_{InE} are balancing coefficients.

In inference phase, the model follows a process similar to CoL, with the difference that $f_{m\to se}^i$ is not transformed into f_{re}^i, but serves as the context for MPU, for full utilization of multimodal features. With features $\{f_{\text{CoL}}^i\}_{i=4}^4$ being input to SegHead, the predicted $\hat{S}_{\text{CoL}}$ serves as the final output.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate our method on three datasets: **DELIVER** (Zhang et al. 2023b) (general scenes with RGB,

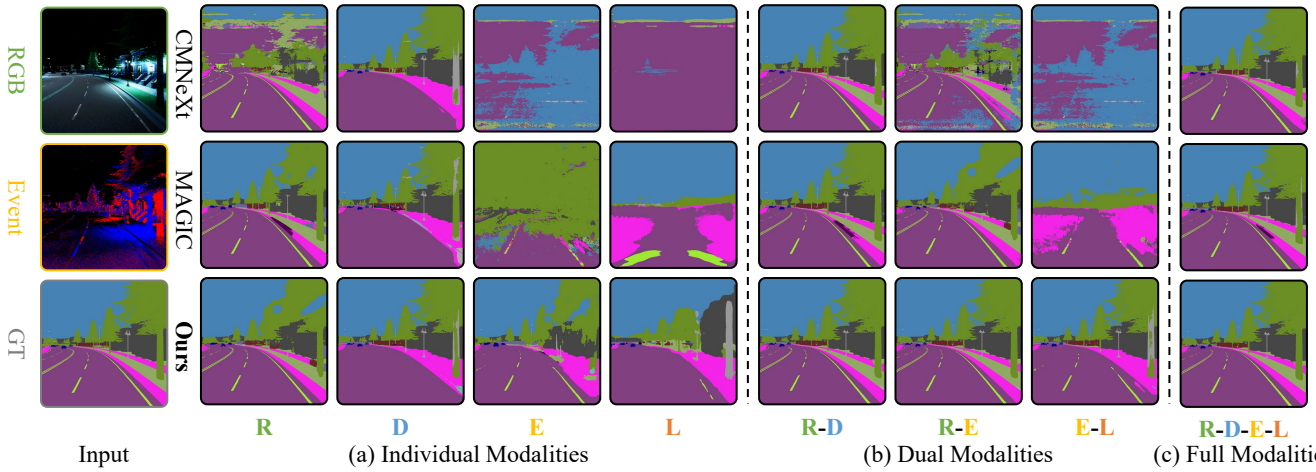

Figure 5: Visual quality comparisons of CMNeXt, MAGIC, and CHARM (Ours) on **DELIVER** dataset of different situations and arbitrary-modal combinations.

Backbones	Methods	DELIVER			MCubeS			MUSES		
		Average↑	*Top-1*↑	*Last-1*↑	*Average*↑	*Top-1*↑	*Last-1*↑	*Average*↑	*Top-1*↑	*Last-1*↑
MiT-B0	CMNeXt	22.05	59.18	0.37	14.35	40.94	0.46	10.80	46.66	2.64
	MAGIC	40.49	63.40	0.26	34.56	47.85	0.55	33.34	49.05	2.68
	Any2Seg	-	-	-	-	-	-	33.86	50.00	3.17
	AnySeg	47.51	59.72	21.74	-	-	-	40.23	51.25	19.57
	CHARM (**Ours**)	**50.97**	**64.03**	**26.47**	**39.05**	**48.24**	**28.37**	**42.19**	**52.31**	**22.31**
	w.r.t. SOTA	*+3.45*	*+0.63*	*+4.73*	*+4.49*	*+0.39*	*+27.82*	*+1.96*	*+1.06*	*+2.74*
Mit-B2	CMNeXt	25.15	66.43	0.72	25.17	51.54	1.54	33.14	58.28	1.23
	MAGIC	44.66	67.66	0.27	38.00	53.01	0.32	36.19	55.36	3.34
	Any2Seg	45.04	68.25	0.31	-	-	-	-	-	-
	CHARM (**Ours**)	**54.96**	**68.43**	**28.76**	**46.58**	**54.33**	**38.50**	**46.18**	**59.36**	**25.28**
	w.r.t. SOTA	*+9.92*	*+0.18*	*+28.04*	*+8.57*	*+1.32*	*+36.96*	*+9.98*	*+1.08*	*+21.93*

Table 1: Objective quality comparison of methods trained with datasets: **DELIVER** (four modalities), **MCubeS** (four modalities), **MUSES** (three modalities). **Bold** and underlined letter the first and second best results. "-" denotes the absence of results due to the unavailability of Any2Seg and AnySeg. All the numbers in the table are the mIoU values (%).

Depth, LiDAR, Event), **MCubeS** (Liang et al. 2022) (material segmentation with RGB, NIR, DoLP, AoLP), and **MUSES** (Brödermann et al. 2024) (driving scenarios with RGB, Event, LiDAR).

Baseline Methods. We compare our method with four advanced methods, *i.e.*, CMNeXt for MSS, and Any2Seg, MAGIC, AnySeg for MaSS. For equality, the backbones of all methods are set to MiT-B0 and MiT-B2 of Segformer (Xie et al. 2021). We also conduct experiments with PVTv2 (Wang et al. 2022b) and Swin Transformer (Liu et al. 2021) in the *supplementary materials*. Note that results for AnySeg and Any2Seg are taken from their reports since there are no released codes or models.

Metrics. We evaluate performance using mean Intersection over Union (mIoU). *Average*, *Top-1*, and *Last-1* denote the average, best, and worst mIoU across all-modal combinations, which assess overall performance, optimal complementary effects, and degree of fragile-modal potential activation, respectively. Detailed mIoU results for all combinations are provided in the *supplementary materials*.

3.2 Qualitative Comparison

Fig. 5(a)-(c) shows the qualitative comparisons of our method with the baseline methods on the **DELIVER** dataset across different modality configurations. Results on other datasets are provided in *supplementary materials*.

(a) **Individual Modalities**: Our method demonstrates consistent performance across all individual modalities by effectively exploiting the inherent potential of each input. It not only achieve better performance on robust modalities of RGB and Depth than the baselines, but also generate good results on fragile modalities of Event and LiDAR, on which the baselines completely fail to handle.

(b) **Dual Modalities**: Our method exhibits superior cross-modal complementarity across various modality combinations. It achieves consistently high performance in robust+robust (R-D) settings. In robust+fragile (R-E) combinations, our method leverages Event information to compensate for RGB limitations, especially in dense tree foliage. Notably, under the fragile+fragile setting (E-L),

Variants	Components			DELIVER			MCubeS			MUSES		
	MPU	CoL	InE	*Average*↑	*Top-1*↑	*Last-1*↑	*Average*↑	*Top-1*↑	*Last-1*↑	*Average*↑	*Top-1*↑	*Last-1*↑
(a)	✗	✗	✗	33.95	63.17	2.30	25.59	45.38	3.39	30.86	51.65	2.26
(b)	✗	✓	✗	41.53	62.50	3.28	33.38	46.83	14.21	31.38	51.39	0.81
(c)	✓	✓	✗	45.23	<u>63.97</u>	15.89	35.22	<u>47.31</u>	20.15	35.88	<u>51.96</u>	15.20
(d)	✗	✓	✓	<u>48.76</u>	62.78	<u>24.40</u>	<u>36.70</u>	46.92	<u>26.38</u>	<u>38.67</u>	51.77	<u>19.97</u>
(e)	✓	✓	✓	**50.97**	**64.03**	**26.47**	**39.05**	**48.24**	**28.37**	**42.19**	**52.31**	**22.31**

Table 2: Ablation study on progressive component integration. The table demonstrates the cumulative benefits of each proposed module: (a) direct addition fusion baseline, (b) addition fusion with CoL, (c) addition fusion with CoL and InE, and (d) complete framework with all components. **Bold** and <u>underlined</u> letter the first and second best results.

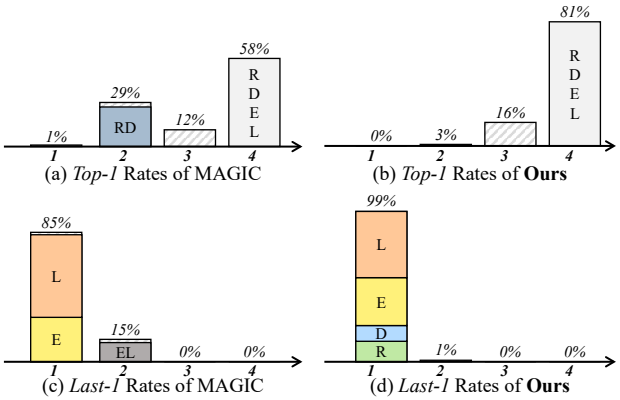

Figure 6: *Top-1* (best performance) and *Last-1* (worst performance) rates across modal combination sizes in **DELIVER**.

our method shows significant improvement in the tree with complementary cues between Event and LiDAR.

(c) **Full Modalities**: When integrating all modalities (R-D-E-L), our method also achieves superior results. While baselines misclassify part of the tree trunk as utility poles, our method correctly identifies it as vegetation, demonstrating the significant complementary effects of our mutual perception mechanism.

3.3 Quantitative Comparison

We conduct quantitative comparisons on all datasets to evaluate the effectiveness of our method across arbitrary modality combinations, as show in Sec. 2.6. It demonstrates that it achieves consistent improvements in *Average*, *Top-1*, and *Last-1* mIoU over all baselines, validating its robustness in handling varying modality configurations. Specifically, on **DELIVER** with MiT-B2, our method achieves 9.92% gain in *Average* mIoU over Any2Seg. Notably, it also delivers notable improvements in handling fragile modalities, with *Last-1* mIoU increasing by over 28.04% across different backbones. These consistent improvements validate the effectiveness of our complementary learning approach in collaborative harmonization across modalities.

Fig. 6 further illustrate insights into modal combination performance by comparing our method with MAGIC. Firstly, the best results as shown in the *Top-1* rates of all modality combinations are predominantly concentrated in combination of all four modalities, while the worst results

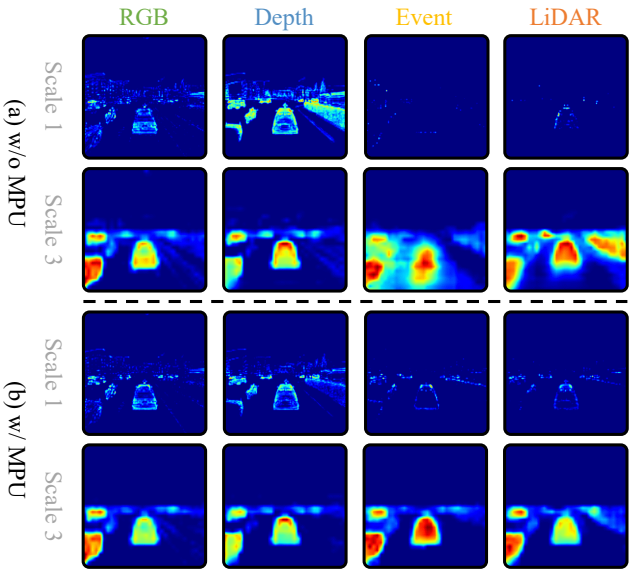

Figure 7: Feature visualization in **Variant (d)** without MPU and **Variant (e)** with MPU.

shown by the *Last-1* rates are more concentrate in the single modality in our method. This validates the effectiveness of the proposed CoL strategy in harnessing and integrating the complementary strengths of each modality, namely extra modality input can ensure better result.

Furthermore, the worst results was most in the fragile modalities of Event and Lidar from MAGIC. In our method, the worst results distribute more evenly in all modalities, owning to the reason that the InE strategy in our method has significantly boost the performance of fragile modalities by stimulating their own advantage and enhancing them through learning from other modalities.

3.4 Ablation Studies

Progressive Component Analysis. The progressive integration of components reveals their distinct functional roles and indicates the superiority of MPU-based methods over simple additive strategies. **Variant (a)** serves as the baseline with direct additive fusion, suffering from severe modality imbalance as evidenced by poor *Last-1* mIoU across all datasets. **Variant (b)** introduces CoL without MPU, showing modest improvements in average performance but still strug-

gling with fragile-modal combinations. Comparing **Variant (c)** with **Variant (b)**, MPU-equipped methods achieve significant gains in *Average* and *Top-1* mIoU, demonstrating MPU's effectiveness in handling fragile-modal scenarios. This superiority stems from MPU's mutual perception mechanism that iteratively inherits complementary content from all modalities, preventing information-rich robust modalities from being diluted by information-sparse fragile modalities as occurs in additive fusion. **Variant (d)** incorporates InE alongside CoL, providing protective mechanism that further alleviate performance suppression of fragile modalities, validated by the improvements in *Last-1* mIoU. **Variant (e)** achieves optimal performance by integrating all components, where MPU harmonizes cross-modal interactions while CoL enables robust complementary fusion and InE ensures individual modal enhancement.

To further validate MPU's capability in collaborative harmonization, we visualize the features by Grad-CAM (Selvaraju et al. 2017) in **Variant (d)** without MPU and **Variant (e)** with MPU in Fig. 7. It shows that at Scale 1, robust modalities (RGB, Depth) show significantly improved on target vehicles with enhanced textural cues extraction, while fragile modalities (Event, LiDAR) show reduced noise and more concentrated responses. At Scale 3, all modalities exhibit substantially refined attention patterns, with Event and LiDAR showing markedly reduced diffusion in irrelevant regions while maintaining their unique semantic contributions. This cross-modal mutual perception mechanism enables each modality to leverage contextual information from others, leading to enhanced discriminative power for robust modalities and focused representational domains for fragile modalities. This validates that MPU enables maximum texture extraction in shallow layers and facilitates semantic exchange without mutual suppression in deeper layers, successfully achieving harmonization rather than homogenization while preserving modality-specific strengths.

Comparison between MPU with Explicit Alignment. To demonstrate the effectiveness of MPU in promoting implicit alignment, we replace the MPU by a explicit alignment that applies KL divergence loss at each scale to force multimodal features toward their average representation. The visualization in Fig. 8 reveals differences in activation patterns: explicit alignment unit forces all modalities toward centralized representations causing universal degradation. MPU maintains distinctive yet complementary activation patterns across modalities. The red boxes show ineffective learning in fragile modalities under explicit constraints, where the model captures features beyond its inherent capacity at Scale 1. The green boxes show degraded activations in robust modalities at Scale 3, with scattered and weakened responses due to forced alignment. In contrast, MPU preserves the representational integrity of each modality across both shallow and deep layers, enabling effective cross-modal cooperation without suppressing modality-specific strengths.

Analysis of Different Sampling Strategies. To validate the fragile modality-biased sampling in InE, we compare three sampling strategies in Tab. 3. *Softmax* emphasizes robust modalities with higher sampling probabilities, *Uniform* samples all modalities equally, and *Softmin* prioritizes fragile modalities. *Softmin* achieves the best performance, improving *Average* mIoU by 2.75%, *Last-1* by 3.11%, and *Top-1* by 0.01%. This validates our design, as CoL already benefits robust modalities through robustness-weighted fusion while InE focuses on enhancing fragile modalities through protected learning spaces, preventing their suppression while maintaining competitive performance on robust modalities.

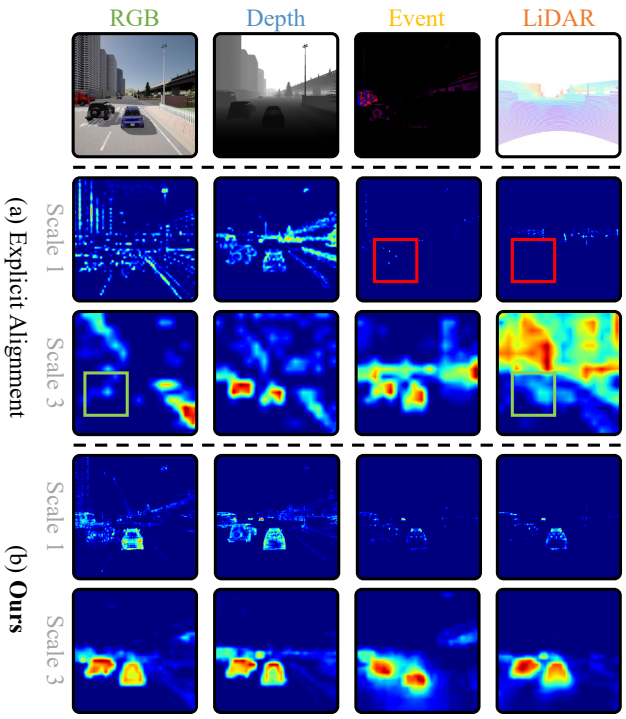

Figure 8: Feature visualization across different strategies. Colored boxes highlight different failure modes.

Strategies	*Average*↑	*Top-1*↑	*Last-1*↑
Softmax	46.80	64.03	16.45
Uniform	48.22	63.88	23.36
Softmin	**50.97**	**64.04**	**26.47**

Table 3: Quality comparison of three different sampling strategies in **DELIVER**.

4 Conclusion

This paper identifies that current MaSS methods pursue multimodal homogenization, leading to suppressed complementary characteristics. To address this, we propose CHARM, a cooperative framework that achieves harmonization through two key innovations: (1) MPU enables discovering modality-interactive correspondences without explicit alignment; (2) CoL and InE strategies balance collaborative learning and individual enhancement through robustness-guided cooperation. Consistent gains across datasets and backbones confirm that CHARM successfully preserves modal distinctiveness and ensures robust MaSS performance across arbitrary modality combinations.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62425102, 62371351), National Key Research and Development Program of Hubei Province (2025BEB011), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

Brödermann, T.; Bruggemann, D.; Sakaridis, C.; Ta, K.; Liagouris, O.; Corkill, J.; and Van Gool, L. 2024. MUSES: The Multi-Sensor Semantic Perception Dataset for Driving Under Uncertainty. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 21–38.

Brödermann, T.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2023. HRFuser: A Multi-Resolution Sensor Fusion Architecture for 2D Object Detection. In *Proceedings of the International Conference on Intelligent Transportation Systems (ITSC)*, 4159–4166.

Dong, S.; Zhou, W.; Xu, C.; and Yan, W. 2024. EGFNet: Edge-Aware Guidance Fusion Network for RGB–Thermal Urban Scene Parsing. *IEEE Transactions on Intelligent Transportation Systems*, 25(1): 657–669.

Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; and Hussain, A. 2023. Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions. *Information Fusion*, 91: 424–444.

Hegde, D.; Yasarla, R.; Cai, H.; Han, S.; Bhattacharyya, A.; Mahajan, S.; Liu, L.; Garrepalli, R.; Patel, V. M.; and Porikli, F. 2025. Distilling Multi-Modal Large Language Models for Autonomous Driving. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 27575–27585.

Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2025. StitchFusion: Weaving Any Visual Modalities to Enhance Multimodal Semantic Segmentation. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1308–1317.

Liang, Y.; Wakaki, R.; Nobuhara, S.; and Nishino, K. 2022. Multimodal Material Segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 19768–19776.

Liao, L.; Chen, W.; Xiao, J.; Wang, Z.; Lin, C.-W.; and Satoh, S. 2022. Unsupervised Foggy Scene Understanding via Self Spatial-Temporal Label Diffusion. *IEEE Transactions on Image Processing*, 31: 3525–3540.

Liao, L.; Chen, W.; Zhang, Z.; Xiao, J.; Yang, Y.; Lin, C.-W.; and Satoh, S. 2023. Only a Few Classes Confusing: Pixel-Wise Candidate Labels Disambiguation for Foggy Scene Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 9992–10002.

Maheshwari, H.; Liu, Y.-C.; and Kira, Z. 2024. Missing Modality Robustness in Semi-Supervised Multi-Modal Semantic Segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 1009–1019.

Reza, M. K.; Prater-Bennette, A.; and Asif, M. S. 2024. MMSFormer: Multimodal Transformer for Material and Semantic Segmentation. *IEEE Open Journal of Signal Processing*, 5: 599–610.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 618–626.

Shi, J.; Shang, C.; Sun, Z.; Yu, L.; Yang, X.; and Yan, Z. 2024. PASSION: Towards Effective Incomplete Multi-Modal Medical Image Segmentation with Imbalanced Missing Rates. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 456–465.

Tan, M.; Zhuang, Z.; Chen, S.; Li, R.; Jia, K.; Wang, Q.; and Li, Y. 2024. EPMF: Efficient Perception-Aware Multi-Sensor Fusion for 3D Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 8258–8273.

Valada, A.; Mohan, R.; and Burgard, W. 2020. Self-Supervised Model Adaptation for Multimodal Semantic Segmentation. *International Journal of Computer Vision*, 128(5): 1239–1285.

Wan, Z.; Zhang, P.; Wang, Y.; Yong, S.; Stepputtis, S.; Sycara, K.; and Xie, Y. 2024. Sigma: Siamese Mamba Network for Multi-Modal Semantic Segmentation.

Wang, H.; Liao, L.; Xiao, J.; Lin, W.; and Wang, M. 2023. Uplink-Assist Downlink Remote-Sensing Image Compression via Historical Referencing. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Wang, M.; Wang, H.; Xiao, J.; and Liao, L. 2022a. A Review of Disentangled Representation Learning for Remote Sensing Data. *CAAI Artificial Intelligence Research*, 1(2): 172–190.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022b. PVT v2: Improved Baselines with Pyramid Vision Transformer. *Computational Visual Media*, 8(3): 415–424.

Wei, S.; Zhou, Z.; Lu, Z.; Yuan, Z.; and Su, B. 2025. HDBFormer: Efficient RGB-d Semantic Segmentation with a Heterogeneous Dual-Branch Framework. *IEEE Signal Processing Letters*, 32: 91–95.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 12077–12090.

Yang, B.; Guo, Y.; Ni, R.; Liu, Y.; Li, G.; and Hu, C. 2025. Asymmetric Multimodal Guidance Fusion Network for Realtime Visible and Thermal Semantic Segmentation. *Engineering Applications of Artificial Intelligence*, 142: 109881.

Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; and Stiefelhagen, R. 2023a. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12): 14679–14694.

Zhang, J.; Liu, R.; Shi, H.; Yang, K.; Reiß, S.; Peng, K.; Fu, H.; Wang, K.; and Stiefelhagen, R. 2023b. Delivering Arbitrary-Modal Semantic Segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1136–1147.

Zhang, Y.; Li, N.; Jiao, J.; Ai, J.; Yan, Z.; Zeng, Y.; Zhang, T.; and Li, Q. 2025. CMFFN: An Efficient Cross-Modal Feature Fusion Network for Semantic Segmentation. *Robotics and Autonomous Systems*, 186: 104900.

Zhang, Z.; Xiao, J.; Liao, L.; and Wang, M. 2024. RefScale: Multi-Temporal Assisted Image Rescaling in Repetitive Observation Scenarios. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 9866–9874.

Zheng, X.; Lyu, Y.; Jiang, L.; Zhou, J.; Wang, L.; and Hu, X. 2024a. MAGIC++: Efficient and Resilient Modality-Agnostic Semantic Segmentation via Hierarchical Modality Selection. arXiv:2412.16876.

Zheng, X.; Lyu, Y.; and Wang, L. 2024. Learning Modality-Agnostic Representation for Semantic Segmentation from Any Modalities. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Proceedings of the European Conference on Computer Vision (ECCV)*, 146–165.

Zheng, X.; Lyu, Y.; Zhou, J.; and Wang, L. 2024b. Centering the Value of Every Modality: Towards Efficient and Resilient Modality-Agnostic Semantic Segmentation. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Proceedings of the European Conference on Computer Vision (ECCV)*, 192–212.

Zheng, X.; Xue, H.; Chen, J.; Yan, Y.; Jiang, L.; Lyu, Y.; Yang, K.; Zhang, L.; and Hu, X. 2025. Learning Robust Anymodal Segmentor with Unimodal and Cross-Modal Distillation. arXiv:2411.17141.