

Article

Mamba-KGSC: Knowledge-Guided Semantic Communication for Robust V2V Cooperative Object Detection

Guangqian Wang *, Jie Sun, Yuqi Liu, Min Huang and Puning Zhang 

School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; s240101123@stu.cqupt.edu.cn (J.S.); 2023210240@stu.cqupt.edu.cn (Y.L.); s250132065@stu.cqupt.edu.cn (M.H.); zhangpn@cqupt.edu.cn (P.Z.)
* Correspondence: 2023210379@stu.cqupt.edu.cn

Abstract

Vehicle-to-Vehicle (V2V) cooperative object detection enhances environmental perception capabilities in complex traffic scenarios by sharing sensory information among vehicles, but limited transmission bandwidth and wireless channel noise can significantly affect the reliable transmission of cross-vehicle semantic features and lead to a degradation in detection performance at the receiver. Although existing semantic communication methods based on DeepJSCC can alleviate the cliff effect of traditional separated source–channel coding under low signal-to-noise ratio conditions, they typically rely on additional external autoencoder structures, which increase model complexity and the deployment burden on vehicular edge computing platforms. Meanwhile, under high compression ratios, these methods struggle to adequately preserve detection-related fine-grained information, such as object boundaries, spatial locations, and local structures. Motivated by these challenges, we develop Mamba-KGSC as a lightweight knowledge-guided semantic communication framework for robust V2V cooperative object detection. At the transmitter, Mamba-KGSC utilizes the internal time-scale parameters of the Mamba-YOLO-T backbone network to generate spatial semantic masks, realizing the sparse encoding and transmission of task-relevant features while avoiding the introduction of complex external codec networks. At the receiver, a multi-source knowledge base constraint verification module is constructed to refine the initial detection results by combining physical consistency screening with visual-physical spatial joint redundancy suppression, thereby suppressing physically inconsistent misdetections and repeated detections induced by channel noise. The experimental evaluation indicates that, under a 50% compression ratio, multiple SNR settings, and different channel models, the front-end semantic communication branch of Mamba-KGSC improves mAP@0.5:0.95 by an average of 1.90 percentage points over the DeepJSCC baseline. The multi-source knowledge base constraint verification module further reduces abnormal and duplicate candidate bounding boxes. Overall, Mamba-KGSC provides a balanced solution in terms of transmission cost, detection accuracy, model complexity, and physical consistency, offering a lightweight implementation scheme for robust V2V cooperative detection in challenging communication environments.



Academic Editor: Fabio Grandi

Received: 2 June 2026

Revised: 24 June 2026

Accepted: 1 July 2026

Published: 3 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: semantic communication; V2V cooperative object detection; Mamba; knowledge-guided; multi-source knowledge base; robust perception

1. Introduction

V2V cooperative object detection can effectively alleviate the issue of incomplete detection in single-vehicle perception caused by limited fields of view, object occlusion, and dis-

tant object degradation by sharing sensory information among vehicles, which has attracted increasing attention as a key topic in cooperative autonomous-driving perception [1,2]. Unlike single-vehicle object detection, V2V cooperative detection not only relies on the accurate recognition of traffic objects by front-end models, but also requires sensory information to be stably transmitted in the Internet of Vehicles (IoV) environment characterized by high-speed mobility, wireless channel perturbations, and limited bandwidth. Directly transmitting raw images or high-dimensional dense features incurs high communication overhead and transmission latency, and increases the computational burden on vehicular edge nodes, making it difficult to meet the real-time, reliability, and lightweight deployment requirements of autonomous-driving systems [3,4]. Recent studies on vehicular edge computing and multimodal integrated sensing and communication (ISAC) further indicate that sensing, communication, and computing resources should be jointly optimized under energy and latency constraints in highly mobile vehicular environments [5]. This observation also motivates the development of lightweight perception–communication frameworks that reduce transmission overhead and deployment burden on vehicular edge platforms.

From the perspective of the information exchange level, existing cooperative perception methods share raw observations, intermediate semantic features, or decision-level detection outputs. Among them, intermediate semantic feature sharing provides a practical trade-off between information richness and communication cost, but it still faces challenges in feature selection, compression, and reliable transmission under bandwidth-limited and noisy V2V links. Therefore, how to transmit task-relevant semantic information efficiently and robustly remains a key issue for improving cooperative object detection in practical IoV scenarios.

Recent studies on intelligent connected systems and multimodal semantic learning have investigated task-aware semantic representation and feature integration from multiple perspectives, including driver emotion recognition with personality cues in intelligent connected-vehicle environments, semantic feature fusion for encrypted retrieval in the Internet of Medical Things, and personality-coupled multimodal sentiment analysis [6–8]. These works show that task-aware multimodal semantic representation and feature integration are valuable for complex perception and information-processing scenarios.

Addressing the issue of sensory information transmission under limited bandwidth, traditional separated source–channel coding (SSCC) typically designs source compression and channel protection separately, possessing a clear modular structure and a relatively mature theoretical foundation [9,10]. However, under low signal-to-noise ratio or fast-changing channel conditions, separated communication links are prone to a sudden performance degradation known as the cliff effect, leading to a significant decline in the recovery quality of images or features at the receiver, which further affects object detection performance. Semantic communication (SemCom) provides a new design paradigm for IoV cooperative perception, where the core objective is no longer the bit-by-bit recovery of the transmitted content, but rather the extraction and transmission of semantic information directly relevant to downstream tasks, in order to improve task execution efficiency under limited communication resources [11].

In recent years, Deep Joint Source–Channel Coding (DeepJSCC) has used end-to-end neural architectures to jointly learn source representations and channel mappings, showing strong robustness under low signal-to-noise ratio (SNR) conditions in image transmission and task-oriented communication scenarios [12–15]. However, existing DeepJSCC-based methods typically rely on independent external autoencoder structures for feature compression and recovery, which lengthens the overall computational pipeline and increases the deployment complexity on vehicular edge platforms. More importantly, external autoencoders tend to uniformly compress the overall features. Under high compression ratios, it

is difficult to explicitly preserve key detection semantics such as foreground object regions, boundary structures, and spatial location information, which may consequently weaken the object localization and category discrimination capabilities at the receiver. Therefore, V2V cooperative object detection requires a lightweight semantic encoding mechanism tightly coupled with the perception backbone network, enabling limited channel resources to be prioritized for the transmission of detection-task-relevant features.

In addition to transmission efficiency and detection accuracy, harsh channels also affect the physical consistency of the detection results at the receiver. Under the influence of wireless noise and feature perturbations, although the semantic features recovered at the receiver can still generate candidate objects, they may contain high-confidence misdetections, targets with unreasonable locations, and repeated detections caused by fragmented local feature responses. This paper refers to such outputs as abnormal detection results induced by communication noise, whose typical manifestations include vehicle candidate boxes falling into non-drivable areas, targets inconsistent with map semantics, and duplicate detection boxes for the same physical object. For safety-critical autonomous-driving tasks, relying solely on data-driven feature reconstruction and detection head outputs is insufficient to fully guarantee the road feasibility, geometric rationality, and scene consistency of candidate objects. Therefore, it is necessary to introduce traffic scene priors at the receiver to perform physical consistency verification and redundancy suppression on the initial detection results.

State Space Models (SSMs) capture long-range dependencies through hidden state recurrences, exhibiting high computational efficiency in long sequence representation learning [16]. Building upon this, Mamba introduces a selective state update and an efficient scanning mechanism, enabling the model to dynamically adjust the information flow based on the input content [17]. In the Mamba-YOLO-T backbone network, the state space modeling mechanism can enhance the global context modeling capability of visual features, and internally generate time-scale parameters related to spatial responses [18]. These parameters reflect the response differences in various spatial locations to task discrimination, serving as crucial clues for identifying foreground object regions and constructing spatial semantic masks. Based on this characteristic, this paper further utilizes the internal time-scale parameters of Mamba-YOLO-T for task-relevant feature screening, realizing lightweight endogenous semantic encoding, thereby avoiding the introduction of complex external codec networks.

The preceding discussion on channel-induced abnormal detections further suggests that receiver-side result refinement should not rely only on learned visual features, but should also exploit traffic-scene priors, such as high-definition (HD) maps, geometric priors, and spatial topological relationships, to verify whether candidate objects satisfy road semantics and scene-level physical consistency [19–21].

Based on the above analysis, this paper proposes a knowledge-guided lightweight semantic communication framework for robust V2V cooperative object detection, named Mamba-KGSC (Mamba-based Knowledge-Guided Semantic Communication). This framework consists of an endogenous semantic encoder at the transmitter, a wireless channel, a semantic decoder at the receiver, and a multi-source knowledge base constraint verification module. At the transmitter, Mamba-KGSC utilizes the internal time-scale parameters of the Mamba-YOLO-T backbone network to generate spatial semantic masks, thereby preserving and sparsely encoding task-relevant features into a lightweight semantic stream for wireless delivery. On the receiver side, the noise-perturbed semantic stream is first decoded and reconstructed to produce initial detection results. The multi-source knowledge base constraint verification module is then employed to perform physical consistency screening and visual–physical spatial joint redundancy suppression on the candidate bounding boxes.

Through this design, Mamba-KGSC reduces transmission overhead while improving the robustness and physical consistency of V2V cooperative object detection under harsh channels. The principal contributions are summarized below:

1. We develop Mamba-KGSC, a lightweight framework that couples knowledge guidance with semantic communication for V2V cooperative object detection. This framework unifies task-relevant semantic encoding at the transmitter, wireless channel transmission, semantic decoding at the receiver, and multi-source knowledge base constraint verification into the same processing pipeline, providing an implementation scheme with lower computational overhead and easier deployment on vehicular edge platforms for cooperative object detection under limited bandwidth and harsh channel conditions.
2. We design an endogenous semantic encoding mechanism based on the internal time-scale parameters of Mamba-YOLO-T. This mechanism leverages the dynamic response information in the state space module to generate spatial semantic masks, selectively suppressing background redundant features and prioritizing limited channel resources for transmitting semantic information related to foreground targets, thereby reducing V2V transmission overhead and enhancing detection stability under noisy channels.
3. We construct a multi-source knowledge base constraint verification mechanism at the receiver. This mechanism integrates static map knowledge, geometric priors, spatial topological constraints, and visual–physical spatial association rules to perform physical consistency screening and joint redundancy suppression on the initial detection candidate bounding boxes, thereby reducing physically inconsistent misdetections and repeated detections caused by channel perturbations.
4. We conduct comprehensive experimental validations for V2V cooperative object detection. By combining the COCO traffic object subset and the OpenCDA/CARLA cooperative perception scenes, this paper verifies the effectiveness of Mamba-KGSC from aspects including detection accuracy, transmission overhead, abnormal bounding box suppression, duplicate detection suppression, and computational complexity, and further analyzes the role of the multi-source knowledge base constraint verification in physical consistency screening and result refinement.

The remainder of this paper is organized as follows. Section 2 reviews the related research work, including V2V cooperative object detection, task-oriented semantic communication, as well as knowledge-driven robust perception and abnormal detection suppression. Section 3 presents the system model of Mamba-KGSC. Section 4 introduces the Mamba-YOLO-T at the transmitter and the endogenous semantic encoding mechanism. Section 5 elaborates on the semantic decoding and the multi-source knowledge base constraint verification method at the receiver. Section 6 reports the experimental evaluation, and Section 7 concludes the paper.

2. Related Work

2.1. V2V-Oriented Cooperative Object Detection

V2V cooperative object detection aims to alleviate the performance degradation in single-vehicle perception caused by occlusion, distant object degradation, and limited fields of view by sharing sensory information among vehicles, thereby improving the completeness of environmental perception and object localization accuracy in complex traffic scenarios [22–24]. Since this improvement is achieved through information exchange among vehicles, existing cooperative perception methods mainly differ in the level of exchanged information, including raw-data-level, feature-level, object-level, and decision-level cooperation. Feature-level methods transmit intermediate representations for cross-

vehicle alignment and fusion, whereas object-level and decision-level methods exchange more compact detection outputs or final decisions with lower communication costs [25,26].

To address cross-vehicle spatial alignment and multi-view fusion in cooperative object detection, existing studies have explored bird's-eye view (BEV) spatial mapping, extrinsic parameter calibration, spatio-temporal synchronization, and multi-scale feature fusion. Relevant methods typically transform the sensory features of neighboring vehicles into a unified reference coordinate system and enhance cross-vehicle information complementarity through attention mechanisms, graph structure modeling, or multi-scale fusion networks [27,28]. In addition, addressing the issue of limited V2V communication resources, some studies have further introduced feature compression, selective sharing, feature pruning, quantization encoding, or communication scheduling strategies to reduce the cross-vehicle transmission burden [29,30]. These works indicate that cooperative object detection depends not only on the visual perception model itself but is also closely related to information selection, communication resource constraints, and cross-vehicle feature fusion mechanisms.

However, most existing V2V cooperative detection methods assume that shared information can be reliably transmitted between vehicles, or merely simplify the modeling by treating communication overhead as a static constraint, giving insufficient consideration to feature degradation and detection anomalies caused by wireless channel noise [31]. In real IoV environments, bandwidth limitations, channel noise, and transmission perturbations may lead to the distortion of shared features, which in turn causes issues such as degraded detection accuracy, cross-category misdetections, repeated detections, and physically inconsistent candidate bounding boxes at the receiver. Therefore, cooperative object detection for practical V2V scenarios still needs to further consider semantic information transmission under channel corruption, detection robustness under noisy channels, and the physical consistency of detection results.

2.2. Task-Oriented Semantic Communication

Semantic communication shifts the communication objective from traditional bit-by-bit recovery to task-oriented semantic information transmission for downstream tasks, providing a new research paradigm for intelligent perception and cooperative decision-making under bandwidth-constrained conditions [32–34]. In this direction, Joint Source–Channel Coding (JSCC) is widely used to simultaneously model the source compression and channel transmission processes. Furthermore, DeepJSCC, as a deep learning-based JSCC scheme, jointly learns source representations and channel mappings through end-to-end neural networks, demonstrating good robustness under low-SNR conditions in scenarios such as image transmission, visual semantic perception, and task-oriented communication [13–15]. Compared with SSCC, DeepJSCC can alleviate the cliff effect under low-SNR conditions to a certain extent, and has thus become an important baseline method in semantic communication research.

Recent object-detection-oriented semantic communication studies have further shown that COCO-style detection benchmarks and COCO-trained detectors can be used to evaluate task-relevant visual semantic transmission. For example, Guo et al. proposed a multi-scale semantic communication framework for object detection and evaluated task-relevant semantic feature transmission on the MS-COCO benchmark under AWGN and Rayleigh fading channels [35]. Safaeipour and Hashemi studied semantic-aware and goal-oriented wireless image transmission for object detection, where a COCO object-detection pre-trained model was used to extract object-level semantics [36]. These studies indicate that COCO-style object detection provides a reasonable evaluation protocol for image-based semantic communication. Nevertheless, they mainly focus on general object-

detection-oriented semantic transmission, whereas the present work further considers V2V cooperative object detection, channel-induced abnormal detections, and receiver-side knowledge-guided physical consistency verification.

Although DeepJSCC and its extensions have achieved good results in noise-robust transmission, they still have certain limitations in V2V cooperative object detection scenarios. First, most methods rely on independent external neural codecs for feature compression and recovery. Although these codecs may improve channel adaptivity, they introduce additional parameters and floating-point operations (FLOPs), thereby increasing the complexity of edge-side deployment [37]. For resource-constrained vehicular edge platforms, the additional codec structure may become a computational bottleneck for real-time cooperative perception systems. Second, external autoencoders usually perform uniform compression and non-linear mapping on the overall features, making it difficult to explicitly distinguish between foreground object regions and background redundant regions. At high compression ratios, fine-grained information relevant to detection, such as object boundaries, spatial relationships, and local structures, may be weakened, thereby affecting the object localization and category discrimination capabilities at the receiver [38].

The aforementioned problems are particularly prominent in the cooperative object detection task. Unlike general image recovery or classification tasks, V2V cooperative detection not only requires the receiver to recover discriminative semantic features, but also needs to maintain the stability of object boundaries, spatial locations, and the geometric structures of candidate bounding boxes. Preserving only the global reconstruction fidelity of features cannot ensure reliable detection outputs. Therefore, there is still a need for a lightweight semantic encoding mechanism tightly coupled with the perception backbone network, enabling limited channel resources to be prioritized for the transmission of task-relevant semantic features while enhancing detection stability under noisy channels and reducing transmission overhead.

2.3. Knowledge-Driven Robust Perception and Abnormal Detection Suppression

In autonomous driving and intelligent perception systems, pure data-driven models are prone to producing anomalous outputs that are inconsistent with the physical constraints of traffic scenes under conditions of distribution shifts, occlusions, noise perturbations, and feature degradation [39]. To enhance the robustness and interpretability of perception systems, existing research has attempted to introduce external knowledge, such as HD maps, geometric priors, spatial topological relationships, knowledge graphs, or rule bases, to impose additional constraints on visual detection results [19,20]. Such knowledge-driven methods can, to a certain extent, compensate for the limitations of pure visual models in physical consistency reasoning, and have been applied to tasks such as drivable-area constraint verification, static obstacle recognition, scene semantic enhancement, and multi-source perception fusion. Recent studies on physics-guided transportation intelligence further indicate that traffic-domain physical rules and spatio-temporal priors can be combined with graph neural networks to enhance traffic-state modeling and prediction reliability [40]. In addition, online HD map construction methods such as HDMaPNet provide structured map priors that can support map-based scene understanding and physical constraint reasoning in autonomous-driving systems [20].

In terms of methodology, knowledge-driven perception methods can be broadly divided into two categories. One category utilizes HD maps, road boundaries, or geometric constraints to filter out anomalous targets projected onto buildings, vegetation, non-road areas, or other non-navigable regions during the post-processing stage. The other category leverages knowledge graphs, scene graph reasoning, or semantic relationship modeling to verify the contextual relationships, spatial topological relationships, and category con-

sistency among candidate targets, thereby enhancing the interpretability and robustness of models in complex traffic scenarios [41]. These studies indicate that external knowledge constraints play an important role in suppressing misdetections, enhancing scene understanding, and ensuring the semantic and logical consistency of perception results.

However, most existing knowledge-driven perception methods are designed for conventional autonomous-driving perception, localization, or planning tasks, and their integration with semantic communication receivers remains insufficient. In particular, under harsh V2V channel conditions, the semantic features decoded at the receiver may simultaneously contain physical location anomalies, category-inconsistent misdetections, and repeated detections caused by fragmented local feature responses, which can all be induced by channel noise. Existing research has not yet fully considered how, after semantic feature reconstruction and initial detection, to jointly apply static map knowledge, geometric priors, spatial topological constraints, and visual–physical spatial association rules for result verification and candidate bounding box refinement. Therefore, V2V semantic communication requires receiver-side knowledge constraint mechanisms that can jointly verify physical consistency and refine candidate bounding boxes after semantic feature reconstruction.

To further clarify the relationship between Mamba-KGSC and existing studies, representative research directions are compared in Table 1, including DeepJSCC-based semantic communication [12–15], task-oriented semantic communication [32–34,38], communication-efficient cooperative perception [25,26,28,30], and knowledge-driven robust perception [19,20,40,41].

Table 1. Qualitative positioning of Mamba-KGSC in relation to representative related research directions.

Comparison Aspect	DJSCC	TO-SC	CE-CP	KD-P	Ours
External codec	✓	△	–	–	–
Backbone semantic mask	–	–	△	–	✓
Noisy-channel robustness	✓	△	△	–	✓
Knowledge verification	–	–	–	✓	✓
V2V detection focus	△	△	✓	△	✓

Note: DJSCC denotes DeepJSCC-based semantic communication; TO-SC denotes task-oriented semantic communication; CE-CP denotes communication-efficient cooperative perception; KD-P denotes knowledge-driven robust perception. ✓ indicates that the corresponding aspect is explicitly considered, △ indicates that it is partially involved or indirectly related, and – indicates that it is not the main focus.

3. Mamba-KGSC System Model

This section presents the system model of Mamba-KGSC and describes its task-oriented semantic communication mechanism from three perspectives: system architecture, end-to-end information transmission process, and optimization objectives. Considering a typical V2V cooperative object detection scenario, the transmitter vehicle first acquires RGB images from onboard cameras and extracts task-relevant visual semantic features through a perception backbone network. Subsequently, the internal time-scale parameters of the Mamba-YOLO-T backbone network are used to generate spatial semantic masks, such that detection-informative regions are retained and organized into a sparse semantic stream for wireless delivery. At the receiver, the channel-corrupted sparse semantic stream is decoded and reconstructed to produce initial detection results, after which the multi-source knowledge base constraint verification module performs physical consistency screening and visual–physical spatial joint redundancy suppression to obtain robust detection outputs.

For readability, the main notations used in the proposed Mamba-KGSC framework are summarized in Table 2.

Table 2. Summary of main notations used in the proposed Mamba-KGSC framework.

Symbol	Description
Feature representation and semantic encoding	
$\mathbf{X}$	Input RGB image captured by the transmitter vehicle.
H_0, W_0	Height and width of the input image.
l, L	Feature scale index and the total number of multi-scale feature layers.
$\mathbf{F}^{(l)}$	Visual semantic feature map at the l -th scale.
C_l, H_l, W_l	Channel number, spatial height, and spatial width of $\mathbf{F}^{(l)}$.
$\Delta_{c,t}$	Channel-specific time-scale parameter at sequence position t .
$\Delta^{(l)}$	Spatial time-scale parameter map at the l -th scale.
$P_{i,j}^{(l)}$	Semantic preservation probability at spatial location (i, j) .
$M_{i,j}^{(l)}$	Binary semantic mask value at spatial location (i, j) .
α, β	Fixed scaling and translation factors in the sigmoid-based probability mapping.
τ_m	Predefined semantic mask threshold used for binarization.
$\Psi^{(l)}$	Sparse semantic feature representation at the l -th scale.
$r_{tx}^{(l)}$	Feature-level transmission ratio at the l -th scale.
Wireless transmission and decoding	
$\mathbf{S}, \hat{\mathbf{S}}$	Packed channel input representation and the channel-corrupted representation at the receiver.
$\mathbf{H}$	Channel fading coefficient matrix.
$\mathbf{Z}$	Additive Gaussian noise term.
$\mathcal{D}(\cdot)$	Semantic decoding and feature reconstruction operation at the receiver.
$\hat{\mathbf{F}}^{(l)}$	Reconstructed feature map at the l -th scale.
$\Omega^{(l)}, \hat{\Omega}^{(l)}$	Retained coordinate set at the transmitter and decoded valid coordinate set at the receiver.
Detection and knowledge-based verification	
$\mathcal{B}_{init}$	Initial candidate bounding box set generated by the detection head.
$\mathcal{B}_{final}$	Final detection result set after knowledge-based verification.
$\mathbf{b}_i$	The i -th candidate bounding box.
$\mathbf{p}_i, \tilde{\mathbf{p}}_i$	3D physical anchor and its 2D ground-plane projection.
ℓ_i	Map semantic label intersected by the semantic ray of candidate $\mathbf{b}_i$.
$\mathcal{K}$	Multi-source knowledge base.
$\mathcal{G}(\cdot; \mathcal{K})$	Knowledge-based constraint verification and result refinement process.
$\mathcal{V}_{lane}, \mathcal{V}_{geo}, \mathcal{V}_{hit}$	Violation indicator functions for lane, geometric, and semantic raycasting constraints.
$R_{vp}(\cdot, \cdot)$	Visual–physical repetition indicator function.
τ_{img}, τ_{phy}	Image-plane and physical-plane distance thresholds for joint redundancy suppression.
Optimization and evaluation	
$\mathcal{L}_{task}$	Object detection task loss.
$R_{sparse}(\Psi)$	Transmission rate of the sparse semantic stream.
τ_{bw}	Upper limit of the available system bandwidth.
Θ_{train}	Trainable parameter set in the semantic encoding and initial detection pipeline.

3.1. System Architecture

As shown in Figure 1, Mamba-KGSC consists mainly of three parts: the transmitter, the wireless channel, and the receiver. Specifically, the transmitter is responsible for extracting and screening task-relevant semantic features from input images. The wireless channel models the channel perturbations, including additive noise and fading effects, during the V2V transmission process. The receiver further includes a semantic decoding and initial detection module, as well as a multi-source knowledge base constraint verification module. Unlike DeepJSCC schemes that rely on external autoencoders, Mamba-KGSC embeds the semantic screening and sparse encoding process within the perception backbone network and introduces a knowledge constraint verification mechanism at the receiver, thereby simultaneously reducing transmission overhead and enhancing the physical consistency of the detection results.

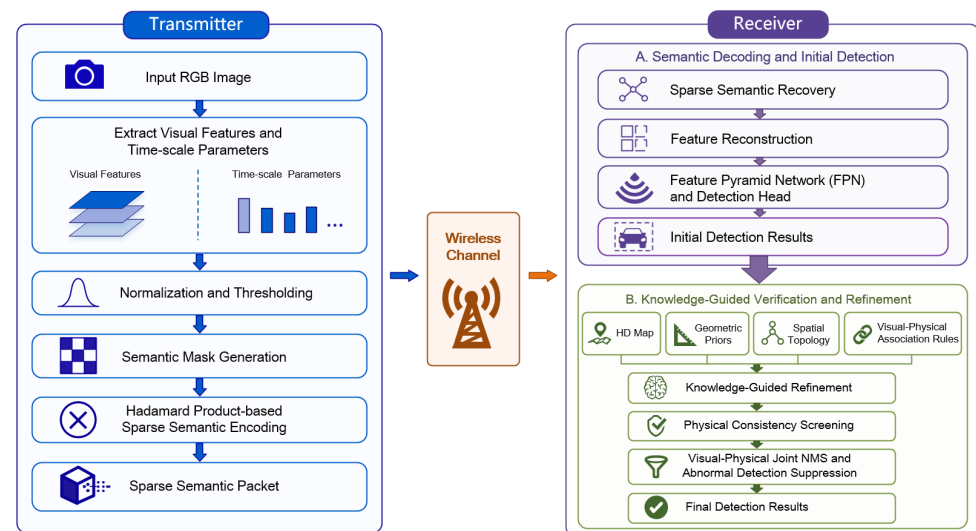


Figure 1. Overall architecture of the Mamba-KGSC system model.

1. **Transmitter-side endogenous semantic encoding.** Given an RGB image captured by the vehicle, this module uses the Mamba-YOLO-T backbone to extract multi-scale visual features together with the associated time-scale parameters. Subsequently, the system generates spatial semantic masks based on the time-scale parameters and screens task-relevant features through element-wise gating operations to obtain a sparse semantic stream for transmission. This process directly reuses the internal response information of the perception backbone network, avoiding the introduction of additional complex external codec networks. The specific implementation of this process is introduced in Section 4.
2. **Wireless channel.** The sparse semantic stream is packed to form the channel input representation and is subjected to channel perturbations during the V2V wireless transmission process. During basic system modeling and noise-aware training, additive white Gaussian noise (AWGN) is adopted to model transmission perturbations under different SNR conditions. To further examine the robustness of the proposed method under fading environments, Rayleigh and Rician fading channels are also considered in the experimental evaluation. This setting is used to simulate the IoV transmission environment under bandwidth limitations, noise perturbations, and typical fading perturbations.
3. **Semantic decoding and initial detection at the receiver.** The receiver first decodes the corrupted sparse semantic representation and reconstructs it into multi-scale visual features. Subsequently, the reconstructed features are fed into the feature pyramid network (FPN) and detection head to generate initial candidate detection results. Since scene-level physical constraints have not yet been incorporated at this stage, the initial results may contain abnormal candidate bounding boxes, cross-category misdetections, and repeated detections induced by channel perturbations.
4. **Multi-source knowledge base.** The multi-source knowledge base module comprehensively utilizes HD maps, geometric priors, spatial topological constraints, and visual–physical spatial association rules to perform physical consistency screening and joint redundancy suppression on the initial candidate bounding boxes. This module adopts a lightweight rule-based verification approach, which is mainly used to eliminate abnormal candidate bounding boxes with unreasonable physical locations and suppress repeated detections caused by channel-induced local feature perturbations. The specific rules are detailed in Section 5.

Through the above architecture, Mamba-KGSC integrates task-relevant semantic encoding, noisy channel transmission, semantic recovery at the receiver, and knowledge constraint verification into a unified processing pipeline, providing a lightweight implementation framework with physical consistency constraints for V2V cooperative object detection under harsh communication conditions.

3.2. V2V Cooperative Object Detection Process

Based on the above system architecture, the end-to-end information transmission process of Mamba-KGSC can be divided into the following four stages.

Step 1: Semantic sparse encoding at the transmitter. Let the RGB image captured by the transmitter vehicle be:

$$\mathbf{X} \in \mathbb{R}^{H_0 \times W_0 \times 3}, \quad (1)$$

where H_0 and W_0 specify the image height and width. Passing $\mathbf{X}$ through the Mamba-YOLO-T backbone yields the following multi-scale visual feature set:

$$\mathcal{F} = \left\{ \mathbf{F}^{(l)} \right\}_{l=1}^L, \quad (2)$$

where L is the number of feature scales, and $\mathbf{F}^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$ is the feature map at the l -th scale. The symbols C_l , H_l , and W_l correspond to the channel number, height, and width of this feature map, respectively.

The transmitter constructs the spatial semantic mask $\mathbf{M}^{(l)}$ based on the time-scale parameters generated internally by the backbone network and performs sparse encoding on the visual features at the l -th scale:

$$\mathbf{\Psi}^{(l)} = \mathbf{F}^{(l)} \odot \mathbf{M}^{(l)}, \quad (3)$$

where $\mathbf{M}^{(l)} \in \{0, 1\}^{H_l \times W_l}$ represents the two-dimensional spatial semantic mask at the l -th scale, and $\odot$ denotes element-wise multiplication. When multiplied by $\mathbf{F}^{(l)}$, $\mathbf{M}^{(l)}$ is broadcast along the channel dimension to match the dimensions of $\mathbf{F}^{(l)}$. $\mathbf{\Psi}^{(l)}$ represents the sparse semantic features at the l -th scale filtered by the semantic mask. The cross-scale sparse semantic representation can be denoted as:

$$\mathbf{\Psi} = \left\{ \mathbf{\Psi}^{(l)} \right\}_{l=1}^L. \quad (4)$$

Step 2: Wireless channel transmission. The sparse semantic representation $\mathbf{\Psi}$ is packed to form the channel input representation to be transmitted:

$$\mathbf{S} = \mathcal{P}(\mathbf{\Psi}), \quad (5)$$

where $\mathcal{P}(\cdot)$ denotes the packing operation of the sparse semantic stream. During transmission, the perturbed representation obtained by the receiver can be written as:

$$\hat{\mathbf{S}} = \mathbf{H} \odot \mathbf{S} + \mathbf{Z}, \quad (6)$$

where $\mathbf{H}$ denotes the channel fading coefficient matrix with the same dimensions as $\mathbf{S}$, $\odot$ denotes element-wise multiplication, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ represents additive white Gaussian noise, $\mathbf{I}$ is the identity matrix, $\hat{\mathbf{S}}$ represents the perturbed semantic representation obtained by the receiver, and σ^2 is the noise variance. For the AWGN channel, $\mathbf{H}$ reduces to an all-one matrix. For the Rayleigh and Rician fading channels, $\mathbf{H}$ is generated according to

the corresponding unit-average-power fading distribution, and the Rician K -factor is set to 2 on a linear scale in the experiments.

Step 3: Semantic decoding and initial detection at the receiver. The receiver decodes the sparse semantic representation from the perturbed representation $\hat{\mathbf{S}}$ and reconstructs the multi-scale feature set:

$$\hat{\Psi} = \mathcal{D}(\hat{\mathbf{S}}), \quad (7)$$

$$\hat{\mathcal{F}} = \left\{ \hat{\mathbf{F}}^{(l)} \right\}_{l=1}^L, \quad (8)$$

where $\mathcal{D}(\cdot)$ denotes the semantic decoding and feature reconstruction operation at the receiver, $\hat{\Psi}$ represents the decoded sparse semantic representation, $\hat{\mathcal{F}}$ represents the reconstructed multi-scale feature set, and $\hat{\mathbf{F}}^{(l)}$ represents the reconstructed feature map at the l -th scale. Subsequently, the reconstructed features are fed into the detection head to generate the initial detection result set:

$$\mathcal{B}_{\text{init}} = \mathcal{H}(\hat{\mathcal{F}}) = \{\mathbf{b}_i\}_{i=1}^{N_{\text{init}}}, \quad (9)$$

where $\mathcal{H}(\cdot)$ denotes the output mapping of the detection head, N_{init} denotes the number of initial candidate bounding boxes, and $\mathbf{b}_i$ represents the i -th candidate bounding box vector.

Step 4: Multi-source knowledge base constraint verification. To improve the physical consistency of the detection results, the initial detection set $\mathcal{B}_{\text{init}}$ is fed into the multi-source knowledge base module. This module combines static map knowledge, geometric priors, spatial topological constraints, and visual–physical spatial association rules to verify and refine the candidate objects, obtaining the final detection results:

$$\mathcal{B}_{\text{final}} = \mathcal{G}(\mathcal{B}_{\text{init}}; \mathcal{K}), \quad (10)$$

where $\mathcal{K}$ represents the multi-source knowledge base, $\mathcal{G}(\cdot; \mathcal{K})$ denotes the constraint verification and result refinement process based on the multi-source knowledge base, and $\mathcal{B}_{\text{final}}$ represents the final detection result set. Through this step, the system can eliminate physically inconsistent abnormal candidate bounding boxes and suppress repeated detections.

The overall process indicates that Mamba-KGSC reduces the transmission overhead through sparse semantic encoding at the transmitter, and enhances the physical consistency and robustness of the detection results through knowledge constraint verification at the receiver. Unlike semantic communication schemes that solely rely on end-to-end feature reconstruction, this framework simultaneously considers task-relevant semantic preservation, robustness to channel noise, and physical constraints in traffic scenes.

3.3. Problem Modeling and Optimization Objectives

From the perspective of task-oriented semantic communication, the objective of Mamba-KGSC is to learn an effective sparse semantic representation for the object detection task under limited transmission resources and noisy channel conditions, and to obtain reliable and physically consistent detection results at the receiver. It should be noted that the perception backbone network, semantic mask generation at the transmitter, and the initial detection pipeline at the receiver contain trainable parameters. By contrast, the multi-source knowledge base constraint verification module mainly serves as a rule-based result refinement process at the receiver and does not participate in gradient optimization. Its role is reflected in performing physical consistency screening and redundancy suppression on the initial candidate bounding boxes.

Let $\mathcal{L}_{\text{task}}$ denote the object detection task loss, $R_{\text{sparse}}(\Psi)$ denote the transmission rate of the sparse semantic stream Ψ , Y_{gt} denote the ground truth annotations, $\mathcal{C}$ denote the channel perturbation process, and Θ_{train} denote the parameter set in the trainable

semantic encoding and initial detection pipeline, for which the optimization objective can be written as:

$$\begin{aligned} \min_{\Theta_{\text{train}}} \quad & \mathbb{E}_{\mathbf{x}, \mathcal{C}} [\mathcal{L}_{\text{task}}(\mathcal{B}_{\text{init}}, Y_{\text{gt}})] \\ \text{s.t.} \quad & R_{\text{sparse}}(\Psi) \leq \tau_{\text{bw}}, \end{aligned} \quad (11)$$

where τ_{bw} represents the upper limit of the available system bandwidth, and $\mathbb{E}_{\mathbf{x}, \mathcal{C}}[\cdot]$ denotes the mathematical expectation over the input image distribution and the channel perturbation process. In the noise-aware training stage, $\mathcal{C}$ is instantiated by AWGN perturbations, while in the fading-channel evaluation, $\mathcal{C}$ further includes Rayleigh or Rician fading coefficients together with additive noise. This objective indicates that, while satisfying the transmission rate constraint, the model needs to learn task-relevant semantic representations that are robust to channel perturbations, thereby improving the accuracy of the initial detection results at the receiver.

After completing the initial detection, the multi-source knowledge base module further performs rule-based constraint verification and result refinement on $\mathcal{B}_{\text{init}}$, obtaining the final detection results:

$$\mathcal{B}_{\text{final}} = \mathcal{G}(\mathcal{B}_{\text{init}}; \mathcal{K}), \quad (12)$$

where $\mathcal{K}$ represents the multi-source knowledge base, and $\mathcal{G}(\cdot; \mathcal{K})$ denotes the knowledge constraint verification process. As a rule-based verification stage at the receiver, this process utilizes static map knowledge, geometric priors, spatial topological constraints, and visual-physical spatial association rules to perform physical consistency screening and joint redundancy suppression on the initial candidate bounding boxes. Therefore, the final detection performance of Mamba-KGSC is evaluated by $\mathcal{B}_{\text{final}}$, while the optimization of the trainable pipeline primarily acts on the transmitter's semantic encoding, channel-robust feature learning, and the initial detection output.

To enhance the detection stability of the model under different channel conditions, this paper further adopts a two-stage noise-aware training strategy. The first stage performs basic detection fine-tuning on Mamba-YOLO-T without injecting channel noise, allowing the model to adapt to the traffic object detection data and input scale settings. The second stage takes the optimal weights from the first stage as initialization, introduces additive white Gaussian noise perturbations during the training process, and employs a curriculum-based SNR sampling strategy for noise-robust fine-tuning. The Rayleigh and Rician fading channels are further used in the experimental evaluation to examine the cross-channel robustness of the trained model without introducing additional channel-specific retraining. The complete training procedure will be presented in Section 5.

4. Endogenous Semantic Encoder

To reduce the V2V semantic transmission overhead without significantly increasing the computational burden on vehicular edge devices, this paper adopts Mamba-YOLO-T as the perception backbone network at the transmitter and constructs an endogenous semantic encoder. Unlike DeepJSCC methods that rely on independent external autoencoders, the proposed encoder does not introduce additional complex compression networks. Instead, it directly utilizes the time-scale parameters generated internally by the state space module of Mamba-YOLO-T to construct spatial semantic masks. This design enables task-relevant region screening to be completed simultaneously with the feature extraction process of the backbone network, allowing limited channel resources to be prioritized for transmitting foreground semantic information related to object detection. Within the overall Mamba-KGSC pipeline, the endogenous semantic encoder is embedded into the backbone feature extraction process rather than appended as an isolated post-processing compression module. It should also be noted that the transmitter-side sparse semantic encoding is

guided by the endogenous time-scale responses of Mamba-YOLO-T, while the multi-source knowledge base is introduced at the receiver side to guide physical consistency verification and candidate-box refinement after semantic decoding, rather than directly updating the transmitter-side encoder.

4.1. Extraction of Endogenous Time-Scale Parameters

Traditional convolutional neural networks mainly rely on local convolutional kernels to model spatial neighborhood relationships. To capture long-range spatial dependencies, it is usually necessary to expand the receptive field through deeper network stacking. Mamba performs state space modeling on visual sequences through a two-dimensional selective scanning mechanism and can obtain global context information with lower computational complexity. After image patch partitioning and linear projection, the input RGB image is represented as a one-dimensional visual sequence, where the sequence position t corresponds to a spatial patch or feature location in the image. The SSM module updates the hidden state using the input feature $\mathbf{x}_t$ at this position and further generates the corresponding output. The overall process of endogenous semantic mask generation and sparse semantic encoding is illustrated in Figure 2.

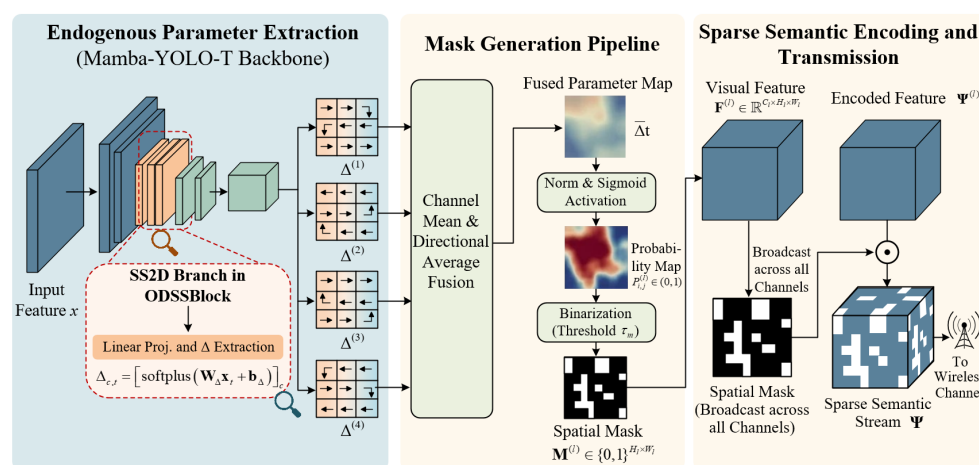


Figure 2. Endogenous semantic mask generation and encoding mechanism based on the Mamba-YOLO-T backbone network.

A key parameter in the Selective Scan 2D (SS2D) module is the time-scale parameter Δ_t , which participates in the discretization process of the continuous state space model and provides a response cue for subsequent spatial semantic mask generation. Specifically, the time-scale parameter $\Delta_{c,t}$ for a specific channel is dynamically generated based on the current input feature:

$$\Delta_{c,t} = [\text{softplus}(\mathbf{W}_\Delta \mathbf{x}_t + \mathbf{b}_\Delta)]_c, \quad (13)$$

where $\mathbf{W}_\Delta$ and $\mathbf{b}_\Delta$ denote the learnable weight matrix and bias vector of the linear projection layer, respectively, $\mathbf{x}_t$ represents the input feature at sequence position t , c denotes the channel index, $[\cdot]_c$ denotes the extraction of the c -th channel component of the vector, and $\text{softplus}(\cdot)$ is used to ensure that the generated time-scale parameter is non-negative. $\Delta_{c,t}$ reflects the state update scale of the state space module across different channels and sequence positions, which can serve as an internal representation of spatial response differences.

In Mamba-YOLO-T, the SS2D mechanism scans the two-dimensional feature representation from $N_{\text{dir}} = 4$ directions to capture non-causal spatial context. To generate a unified two-dimensional spatial response map, the time-scale parameters from the four scanning branches are aggregated uniformly. First, the time-scale parameters are averaged along the channel dimension to obtain a scalar response for each sequence position. Subsequently,

an element-wise average is performed across different scanning directions to mitigate directional bias:

$$\bar{\Delta}_t = \frac{1}{N_{\text{dir}}} \sum_{k=1}^{N_{\text{dir}}} \left(\frac{1}{C_{\Delta}} \sum_{c=1}^{C_{\Delta}} \Delta_{c,t}^{(k)} \right), \quad (14)$$

where C_{Δ} is the channel dimension of the parameter tensor, and $\Delta_{c,t}^{(k)}$ is the parameter generated by the k -th scanning branch at sequence position t . The directionally averaged parameter $\bar{\Delta}_t$ is then reshaped back into a two-dimensional spatial format. Therefore, at each feature pyramid scale l , the model obtains a pair of scale-specific representations, including the visual semantic feature map $\mathbf{F}^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$ and the corresponding spatial time-scale parameter map $\Delta^{(l)} \in \mathbb{R}^{H_l \times W_l}$. Here, C_l is the channel number of $\mathbf{F}^{(l)}$, while H_l and W_l specify the shared spatial height and width of $\mathbf{F}^{(l)}$ and $\Delta^{(l)}$.

4.2. Semantic Mask Generation and Encoding

Because the dynamic range of the time-scale response maps $\Delta^{(l)}$ varies significantly across different network layers, directly applying a threshold to the raw time-scale response maps would lead to severe feature imbalance. Therefore, spatial statistics-based normalization is first performed on $\Delta^{(l)}$, mapping it into a continuous probability space:

$$P_{i,j}^{(l)} = \text{Sigmoid} \left(\alpha \frac{\Delta_{i,j}^{(l)} - \mu(\Delta^{(l)})}{\sigma(\Delta^{(l)})} + \beta \right), \quad (15)$$

where $\Delta_{i,j}^{(l)}$ is the scalar value of the parameter map at spatial location (i, j) at scale l , i denotes the height index, j denotes the width index, $\text{Sigmoid}(\cdot)$ is used to map the normalized parameters into the $(0, 1)$ interval, $\mu(\cdot)$ and $\sigma(\cdot)$ are the spatial mean and standard deviation of the parameter map $\Delta^{(l)}$, α and β are the scaling and translation factors, respectively, and ϵ is a small constant used to prevent division-by-zero errors. In the implementation, α and β are treated as fixed hyperparameters rather than learnable parameters. The scaling factor α controls the steepness of the sigmoid mapping and adjusts the contrast between high-response and low-response regions, while the translation factor β controls the global offset of the preservation probability. In this work, β is set to zero to avoid introducing an additional global bias, and α is selected as a moderate fixed value to enhance semantic contrast without making the mask overly sensitive to small response fluctuations. The semantic mask threshold τ_m is used to control the sparsity of the binary mask and thus determines the corresponding feature-level transmission ratio and compression ratio. The obtained $P_{i,j}^{(l)} \in (0, 1)$ represents the semantic preservation probability response at spatial location (i, j) , which is used to measure the likelihood of the feature at that location being preserved for transmission.

To reduce transmission overhead, this paper sets a predefined semantic threshold τ_m to binarize the probability response map into a two-dimensional spatial semantic mask:

$$M_{i,j}^{(l)} = \mathbb{I}(P_{i,j}^{(l)} \geq \tau_m), \quad (16)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. When $M_{i,j}^{(l)} = 1$, it indicates that the feature at spatial location (i, j) is preserved. When $M_{i,j}^{(l)} = 0$, it indicates that the background redundant feature at this location will be suppressed prior to transmission. During the two-stage noise-aware training, α , β , and τ_m are kept fixed and are not updated by backpropagation. This design keeps the compression ratio and communication overhead controllable across different SNR conditions, while the backbone parameters of Mamba-YOLO-T are optimized to learn more noise-robust feature representations. A larger τ_m preserves fewer spatial

feature vectors and reduces communication overhead, whereas a smaller τ_m retains more features and may better preserve detection details at the cost of higher transmission load.

Subsequently, sparse semantic encoding is realized through the Hadamard product between the visual feature map and the spatial semantic mask:

$$\Psi_{c,i,j}^{(l)} = F_{c,i,j}^{(l)} \cdot M_{i,j}^{(l)}, \quad (17)$$

where $c \in \{1, \dots, C_l\}$ denotes the channel index of the visual feature map at the l -th scale, $F_{c,i,j}^{(l)}$ represents the original visual feature of channel c at spatial location (i, j) , and $M_{i,j}^{(l)}$ represents the binary semantic mask at the corresponding location. This operation is equivalent to broadcasting the two-dimensional spatial semantic mask across all C_l channels. Through this gating mechanism, background redundant features at locations with low semantic responses are suppressed before transmission, thereby reducing irrelevant information entering the semantic stream.

Finally, the system adopts a sparse storage strategy, packing only the non-zero feature vectors and their two-dimensional spatial coordinate indices (i, j) into the sparse semantic stream for transmission. The sparse semantic representations across all scales are uniformly denoted as:

$$\Psi = \left\{ \Psi^{(l)} \right\}_{l=1}^L, \quad (18)$$

where L denotes the number of multi-scale feature layers, and $\Psi^{(l)}$ represents the sparse semantic feature representation at the l -th scale filtered by the semantic mask. This definition also provides a unified input representation for the semantic decoding and feature reconstruction process at the receiver in Section 5. Compared with transmitting the dense feature map $F^{(l)}$, the proposed sparse semantic encoding only transmits the spatial feature vectors whose mask value is equal to one. Therefore, the feature-level transmission ratio at the l -th scale can be expressed as:

$$r_{\text{tx}}^{(l)} = \frac{1}{H_l W_l} \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} M_{i,j}^{(l)}, \quad (19)$$

and the corresponding feature-level communication saving ratio is $1 - r_{\text{tx}}^{(l)}$. This formulation explicitly shows that the communication overhead is reduced by decreasing the number of transmitted spatial feature vectors rather than uniformly compressing all feature responses.

The aforementioned process can be viewed as a lightweight semantic source screening mechanism. The time-scale parameters are used to generate spatial semantic masks to selectively preserve task-relevant foreground features and suppress background redundancy. Therefore, the reduction in communication overhead does not come from random feature dropping or uniform feature compression, but from the selective transmission of spatial locations that are more relevant to foreground objects, object boundaries, and localization cues. Meanwhile, the noise-aware curriculum training enables the backbone network to learn feature representations that are more stable against wireless channel noise through progressive perturbations from high-SNR to low-SNR conditions. Compared with DeepJSCC methods that rely on external autoencoders for feature compression, this mechanism directly reuses the time-scale parameters inside the perception backbone network to complete semantic feature screening and sparse encoding, avoiding reliance on complex external compression networks. As a result, Mamba-KGSC can reduce feature-level communication overhead while preserving detection-related semantic information, which helps maintain object detection accuracy under bandwidth-constrained and noisy channel conditions.

5. Semantic Decoding and Multi-Source Knowledge Base Constraint Verification

In a noisy V2V communication environment, the sparse semantic stream obtained at the receiver is subjected to channel perturbations, leading to local degradation of the reconstructed features. Although the detection head can still output candidate targets, the initial detection results may contain physically inconsistent misdetections, cross-category misdetections, and repeated candidate bounding boxes induced by channel noise. To improve the reliability of the detection results at the receiver, this paper constructs a receiver processing pipeline consisting of semantic decoding and multi-source knowledge base constraint verification. In this pipeline, the semantic decoder recovers the perturbed sparse semantic stream into multi-scale features usable by the detection head. The multi-source knowledge base constraint verification module further integrates static map knowledge, geometric priors, spatial topological constraints, and visual–physical spatial association rules to perform physical consistency screening and result refinement on the initial detection results.

5.1. Semantic Decoder

At the l -th scale, the sparse semantic data generated at the transmitter are defined as:

$$\Psi^{(l)} = \left\{ \left((i, j), \psi_{i,j}^{(l)} \right) \mid (i, j) \in \Omega^{(l)} \right\}, \quad (20)$$

where $\Omega^{(l)}$ is the valid coordinate set retained at scale l , and $\psi_{i,j}^{(l)} \in \mathbb{R}^{C_l}$ is the non-zero semantic vector corresponding to location (i, j) . After passing through the wireless channel, the receiver obtains the following channel-perturbed sparse semantic representation:

$$\hat{\Psi}^{(l)} = \left\{ \left((i, j), \hat{\psi}_{i,j}^{(l)} \right) \mid (i, j) \in \hat{\Omega}^{(l)} \right\}, \quad (21)$$

where $\hat{\Omega}^{(l)}$ indicates the coordinate set that is successfully decoded at the receiver, while $\hat{\psi}_{i,j}^{(l)}$ denotes the corresponding semantic vector after channel perturbation.

At scale l , the reconstructed feature map is denoted by $\hat{\mathbf{F}}^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$, where C_l , H_l , and W_l correspond to the channel number, spatial height, and spatial width. The reconstruction rule is written as:

$$\hat{\mathbf{F}}^{(l)}(:, i, j) = \begin{cases} \hat{\psi}_{i,j}^{(l)}, & (i, j) \in \hat{\Omega}^{(l)}, \\ \mathbf{0}_{C_l}, & (i, j) \notin \hat{\Omega}^{(l)}, \end{cases} \quad (22)$$

where $\hat{\mathbf{F}}^{(l)}(:, i, j)$ refers to the reconstructed channel vector at spatial position (i, j) at scale l , $\hat{\Omega}^{(l)}$ is the decoded valid index set, $\hat{\psi}_{i,j}^{(l)}$ is the received semantic vector after noise perturbation, and $\mathbf{0}_{C_l}$ denotes a C_l -dimensional zero vector. This rule places the received sparse features back into their original spatial positions and fills the missing background positions with zeros, producing complete reconstructed feature maps across scales. After performing the aforementioned feature reconstruction for all scales, the receiver can obtain the reconstructed multi-scale feature set:

$$\hat{\mathcal{F}} = \left\{ \hat{\mathbf{F}}^{(l)} \right\}_{l=1}^L, \quad (23)$$

where L denotes the number of multi-scale feature layers, and $\hat{\mathbf{F}}^{(l)}$ represents the reconstructed feature map recovered at the l -th scale. Subsequently, the reconstructed semantic features are fed into the feature pyramid network and detection head to generate the initial detection result set:

$$\mathcal{B}_{\text{init}} = \mathcal{H}(\hat{\mathcal{F}}) = \{\mathbf{b}_i\}_{i=1}^{N_{\text{init}}}, \quad (24)$$

where N_{init} denotes the number of initial candidate bounding boxes, and a single candidate bounding box is denoted as:

$$\mathbf{b}_i = (x_i, y_i, w_i, h_i, c_i, s_i), \quad (25)$$

where (x_i, y_i) are the center coordinates of the candidate bounding box, w_i and h_i denote the width and height, respectively, c_i is the predicted category, and s_i is the corresponding confidence score. This stage focuses on generating initial candidate detection results from the reconstructed features but does not yet incorporate scene-level physical constraints. Therefore, $\mathcal{B}_{\text{init}}$ may still contain abnormal candidate bounding boxes and repeated detection results.

5.2. Multi-Source Knowledge Base

Because monocular RGB detection results themselves do not contain explicit depth information, this paper utilizes camera poses, vehicle poses, and semantic map information in the cooperative driving simulation environment to perform semantic raycasting for each 2D candidate bounding box. Specifically, the system takes the bottom center of the candidate bounding box as the image-domain reference anchor and generates a semantic ray along the projection direction based on the camera imaging model and scene coordinate transformation relationships. It then obtains the corresponding physical anchor vector $\mathbf{p}_i$ and the map semantic label ℓ_i intersected by the semantic ray in the map space. This approach avoids dense retrieval of large-scale HD map polygons, while providing a unified physical reference for subsequent road semantic constraints, geometric constraints, and visual–physical spatial joint redundancy suppression.

To improve the physical consistency and scene consistency of the detection results at the receiver, this paper introduces a multi-source knowledge base module on top of the initial detection results. Unlike a single visual post-processing method, this module integrates multi-source heterogeneous prior information to perform physical consistency verification and result refinement on candidate targets. This paper defines the multi-source knowledge base as:

$$\mathcal{K} = \{\mathcal{K}_{\text{map}}, \mathcal{K}_{\text{geo}}, \mathcal{K}_{\text{topo}}, \mathcal{K}_{\text{assoc}}\}, \quad (26)$$

where $\mathcal{K}_{\text{map}}$ denotes the static scene knowledge provided by the HD map, $\mathcal{K}_{\text{geo}}$ denotes the category-related geometric prior knowledge, $\mathcal{K}_{\text{topo}}$ denotes the spatial topological constraints between the target and the road space, and $\mathcal{K}_{\text{assoc}}$ denotes the visual–physical spatial association rules used to suppress repeated detections.

For any initial candidate bounding box $\mathbf{b}_i \in \mathcal{B}_{\text{init}}$, the system first maps it to the global physical space to obtain its corresponding 3D physical anchor vector:

$$\mathbf{p}_i = \Pi(\mathbf{b}_i) = (X_i, Y_i, Z_i)^T, \quad (27)$$

where $\Pi(\cdot)$ denotes the spatial projection mapping jointly determined by camera parameters, vehicle pose information, and scene coordinate transformations, $\mathbf{p}_i \in \mathbb{R}^3$ represents the 3D physical anchor of the candidate bounding box $\mathbf{b}_i$ in the global physical coordinate system, and X_i , Y_i , and Z_i represent its 3D coordinate components, respectively. Based on this anchor, the system further performs a semantic query based on the static map knowledge to obtain its corresponding map semantic label:

$$\ell_i = \phi(\mathbf{p}_i; \mathcal{K}_{\text{map}}), \quad (28)$$

where $\phi(\cdot)$ denotes the semantic query operator based on spatial indexing, and ℓ_i represents the scene semantic label corresponding to the anchor $\mathbf{p}_i$, such as a drivable area, vegetation, pole, building, or pedestrian area.

Based on this, the multi-source knowledge base does not merely rely on visual confidence for screening, but jointly evaluates the spatial location, geometric morphology, and semantic consistency of the candidate targets through the joint modeling of $\mathbf{b}_i$, $\mathbf{p}_i$, and ℓ_i , thereby providing unified knowledge support for subsequent rule verification and result refinement. The overall processing pipeline of the multi-source knowledge base for the initial detection results at the receiver is shown in Figure 3. This pipeline first completes feature reconstruction and initial candidate bounding box generation from the corrupted semantic stream, and subsequently integrates static map knowledge, geometric priors, and spatial topological constraints to perform physical consistency verification and visual–physical spatial joint redundancy suppression on the candidate targets, ultimately outputting the refined detection results.

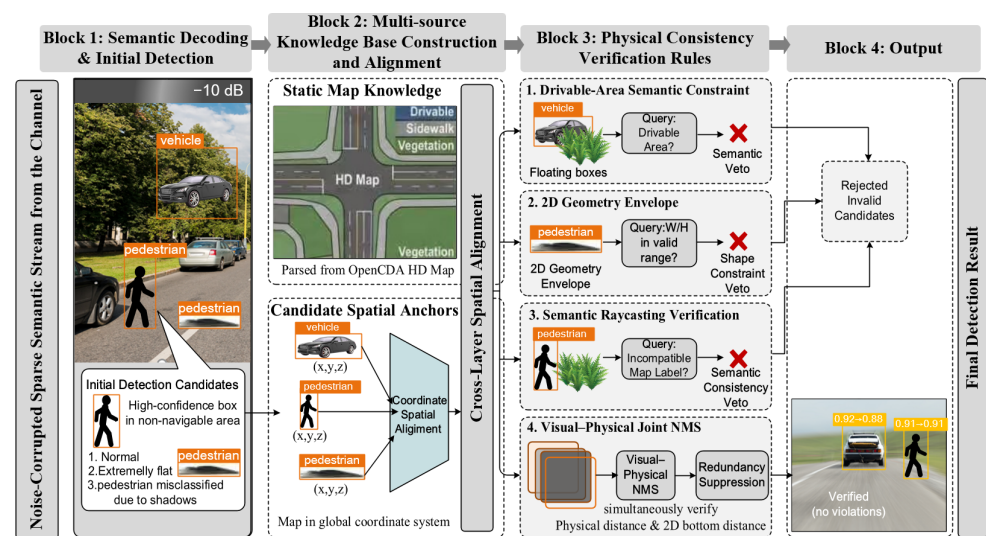


Figure 3. Multi-source knowledge base-driven result verification pipeline at the receiver.

5.3. Physical Consistency Verification Rules

To improve the physical consistency of the initial candidate detection results, this paper formalizes the multi-source knowledge base constraint verification into the following four categories of rules. First, the vehicle category set is defined as $\mathcal{C}_{\text{veh}}$, the drivable area set in the map is defined as $\mathcal{M}_{\text{drive}}$, and the static obstacle semantic set incompatible with the target category c_i is defined as $\mathcal{C}_{\text{forbid}}^{(c_i)}$. Based on this, physical consistency verification is performed on each candidate bounding box $\mathbf{b}_i = (x_i, y_i, w_i, h_i, c_i, s_i)$, where (x_i, y_i) denotes the center coordinates of the candidate bounding box, w_i and h_i denote the width and height, respectively, c_i denotes the predicted category, and s_i denotes the detection confidence.

(1) Drivable-area semantic constraint. For vehicle-class targets, if their projected anchor does not lie within the drivable area, they are considered to violate the road semantic constraint. The corresponding violation indicator function is defined as:

$$\mathcal{V}_{\text{lane}}(\mathbf{b}_i) = \begin{cases} 1, & c_i \in \mathcal{C}_{\text{veh}} \text{ and } \tilde{\mathbf{p}}_i \notin \mathcal{M}_{\text{drive}}, \\ 0, & \text{otherwise,} \end{cases} \quad (29)$$

where $\mathcal{C}_{\text{veh}}$ denotes the vehicle category set, $\mathcal{M}_{\text{drive}}$ denotes the drivable area set in the map, and $\tilde{\mathbf{p}}_i = (X_i, Y_i)^T$ represents the 2D physical anchor position on the global ground plane after the semantic raycasting of candidate bounding box $\mathbf{b}_i$. When $c_i \in \mathcal{C}_{\text{veh}}$ and $\tilde{\mathbf{p}}_i \notin \mathcal{M}_{\text{drive}}$, the candidate bounding box is determined to violate the road semantic constraint, eliminating abnormal vehicle-class candidate bounding boxes appearing in

non-road areas, such as floating vehicles or vehicle candidates erroneously falling inside building areas.

(2) 2D geometric envelope constraint. To suppress extremely deformed candidate bounding boxes caused by noise perturbations, the aspect ratio of the candidate bounding box is defined as:

$$\gamma_i = \frac{w_i}{h_i}. \quad (30)$$

Let the reasonable aspect ratio interval corresponding to category c_i be $[\gamma_{\min}^{(c_i)}, \gamma_{\max}^{(c_i)}]$. Then the geometric violation indicator function can be written as:

$$\mathcal{V}_{\text{geo}}(\mathbf{b}_i) = \mathbb{I}(\gamma_i \notin [\gamma_{\min}^{(c_i)}, \gamma_{\max}^{(c_i)}]), \quad (31)$$

where w_i and h_i denote the width and height of the candidate bounding box, respectively, and $\mathbb{I}(\cdot)$ is the indicator function. This constraint is used to identify abnormal detection results that clearly do not conform to the shape distribution of real objects, such as extremely flat pedestrians or extremely elongated vehicles.

(3) Semantic raycasting constraint. For each candidate bounding box, its physical anchor $\mathbf{p}_i$ corresponds to a map semantic label ℓ_i intersected by the semantic ray. If this label belongs to the static obstacle semantic set incompatible with the candidate category c_i , the candidate target is considered to violate scene semantic consistency. The corresponding violation indicator function is defined as:

$$\mathcal{V}_{\text{hit}}(\mathbf{b}_i) = \mathbb{I}(\ell_i \in \mathcal{C}_{\text{forbid}}^{(c_i)}), \quad (32)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\mathcal{C}_{\text{forbid}}^{(c_i)}$ denotes the static obstacle semantic set incompatible with the candidate category c_i . When $\ell_i \in \mathcal{C}_{\text{forbid}}^{(c_i)}$, the candidate bounding box is determined to have an inconsistency between its category and the scene semantics. This constraint verifies the category and scene semantic consistency of the candidate target by establishing the association among the candidate bounding box, physical anchor, and map semantic label. For example, when a candidate bounding box is predicted as a pedestrian, but the map semantic label intersected by its semantic ray is a static obstacle such as a pole, this candidate result is determined to be inconsistent in category and scene semantics.

(4) Visual–physical spatial joint non-maximum suppression (NMS) constraint. Under severe channel noise perturbations, the local feature responses of the same real target may become fragmented, causing the detection head to generate multiple repeated candidate bounding boxes at similar image locations or similar physical anchors. To this end, this paper defines the visual–physical spatial repetition determination function as:

$$R_{\text{vp}}(\mathbf{b}_i, \mathbf{b}_j) = \mathbb{I}(\|\mathbf{q}_i - \mathbf{q}_j\|_2 < \tau_{\text{img}} \text{ or } \|\tilde{\mathbf{p}}_i - \tilde{\mathbf{p}}_j\|_2 < \tau_{\text{phy}}), \quad (33)$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathbf{q}_i = (u_i, v_i)^T$ and $\mathbf{q}_j = (u_j, v_j)^T$ denote the bottom center anchors of candidate bounding boxes $\mathbf{b}_i$ and $\mathbf{b}_j$ in the image plane. $\tilde{\mathbf{p}}_i = (X_i, Y_i)^T$ and $\tilde{\mathbf{p}}_j = (X_j, Y_j)^T$ denote the physical anchor positions on the global ground plane after the semantic raycasting of candidate bounding boxes $\mathbf{b}_i$ and $\mathbf{b}_j$. $\|\cdot\|_2$ denotes the Euclidean norm. τ_{img} and τ_{phy} denote the image plane distance threshold and physical plane distance threshold, respectively. If $R_{\text{vp}}(\mathbf{b}_i, \mathbf{b}_j) = 1$, it indicates that the two candidate bounding boxes are close enough in either the visual domain or physical space, and can be regarded as repeated candidates generated by the same real target or the same local noise response. The main parameter settings used in the multi-source knowledge base constraint verification module are summarized in Table 3. These parameters are fixed during the

experiments and are used for physical consistency screening and visual–physical spatial redundancy suppression.

Table 3. Main parameter settings of the multi-source knowledge base constraint verification module.

Parameter	Setting
Lane-semantic constraint	Enabled
2D geometric prior constraint	Enabled
Semantic raycasting constraint	Enabled
Visual–physical joint NMS	Enabled
Image-plane NMS distance threshold τ_{img}	15 pixels
Physical-plane NMS distance threshold τ_{phy}	1.5 m
Pedestrian aspect-ratio interval	[0.15, 0.65]
Default vehicle size prior	4.7 m $\times$ 1.9 m $\times$ 1.5 m
OpenCDA/CARLA town	Town06
Static map query radius	50 m

Synthesizing the above rules, the candidate set after physical consistency screening can first be defined as:

$$\mathcal{B}_{\text{valid}} = \{\mathbf{b}_i \in \mathcal{B}_{\text{init}} \mid \mathcal{V}_{\text{lane}}(\mathbf{b}_i) + \mathcal{V}_{\text{geo}}(\mathbf{b}_i) + \mathcal{V}_{\text{hit}}(\mathbf{b}_i) = 0\}, \quad (34)$$

where $\mathcal{B}_{\text{init}}$ denotes the initial candidate bounding box set generated by the detection head at the receiver, and $\mathcal{B}_{\text{valid}}$ denotes the candidate set filtered by the lane semantic, geometric envelope, and semantic raycasting constraints. This set only preserves the candidate bounding boxes that do not violate the physical consistency rules. Subsequently, visual–physical spatial joint non-maximum suppression is performed on $\mathcal{B}_{\text{valid}}$ to obtain the final detection result set:

$$\mathcal{B}_{\text{final}} = \{\mathbf{b}_i \in \mathcal{B}_{\text{valid}} \mid \nexists \mathbf{b}_j \in \mathcal{B}_{\text{valid}}, j \neq i, R_{\text{vp}}(\mathbf{b}_i, \mathbf{b}_j) = 1 \text{ and } s_j > s_i\}. \quad (35)$$

This formula indicates that if a candidate bounding box $\mathbf{b}_i$ and another candidate bounding box $\mathbf{b}_j$ with a higher confidence score are determined to be repeated candidates, $\mathbf{b}_i$ will be suppressed. Otherwise, $\mathbf{b}_i$ is preserved in the final detection set.

Through the above constraints, the multi-source knowledge base constraint verification module at the receiver can perform physical consistency screening and repeated candidate suppression on the initial detection results, thereby improving the stability of the final detection results under channel perturbation conditions. The fixed parameter settings listed in Table 3 are used consistently in the following ablation studies and visualization analysis.

5.4. Two-Stage Noise-Aware Training Process

After completing the definitions of the endogenous semantic encoding at the transmitter, wireless channel perturbation modeling, semantic decoding at the receiver, and multi-source knowledge base constraint verification mechanism, this paper further provides the overall training pipeline of Mamba-KGSC, as shown in Algorithm 1. To ensure that the model possesses high-precision basic perception while effectively resisting channel perturbations in complex communication environments and extracting robust semantic representations, this paper designs a progressive two-stage training strategy.

The first stage is basic perception fine-tuning without channel noise. The model first loads the initial pre-trained weights Θ_0 and conducts E_1 epochs of training in an ideal channel environment without communication noise. For each data mini-batch $(\mathbf{X}, Y_{\text{gt}})$ sampled from the training set $\mathcal{D}$, the model performs forward propagation to generate initial detection results $\mathcal{B}_{\text{init}}$, and back-propagates to update the network parameters Θ_{train}

by calculating the object detection loss $\mathcal{L}_{\text{task}}(\mathcal{B}_{\text{init}}, Y_{\text{gt}})$. The core purpose of the first stage is to allow the backbone network to fully adapt to the scale and data distribution of the traffic-object data, establishing stable feature representations. After this stage ends, the system retains the model weights with the optimal validation accuracy as the initialization parameters for the next stage.

The second stage is noise-aware curriculum fine-tuning. After loading the optimal weights obtained in the first stage, the system introduces wireless channel perturbations during the feature transmission process. In the noise-robust fine-tuning stage, this paper adopts a curriculum-based SNR sampling strategy. Let e denote the current training epoch of the second-stage noise-aware fine-tuning, where $e = 1, \dots, E_2$. The lower SNR sampling bound at the e -th epoch is denoted as $\gamma_{\min}(e)$, which gradually decreases from the initial lower bound $\gamma_{\text{start}} = 5$ dB to the final lower bound $\gamma_{\text{end}} = -10$ dB according to the training progress, while the upper sampling bound is fixed at $\gamma_{\max} = 30$ dB. Subsequently, the training signal-to-noise ratio γ for the current batch is randomly sampled from the interval $[\gamma_{\min}(e), \gamma_{\max}]$. For a given compression ratio setting, the semantic mask threshold τ_m is kept fixed during the training and testing phases. Experiments with different compression ratios are realized by adjusting τ_m . In the experiments, all input images are resized to $640 \times 640 \times 3$. The first-stage clean fine-tuning is conducted for $E_1 = 180$ epochs, and the second-stage noise-aware curriculum fine-tuning is conducted for $E_2 = 120$ epochs. The model is optimized using AdamW with an initial learning rate of 1×10^{-4} , a final learning-rate factor of 0.01, a batch size of 8, and mixed-precision training.

Algorithm 1 Two-stage noise-aware training procedure for Mamba-KGSC

Input: Training dataset $\mathcal{D}$, initial weights Θ_0 , number of basic fine-tuning epochs E_1 , number of noise-aware fine-tuning epochs E_2 , semantic mask threshold τ_m , with $E_1 = 180$ and $E_2 = 120$ in the implementation, initial lower SNR bound $\gamma_{\text{start}} = 5$ dB, final lower SNR bound $\gamma_{\text{end}} = -10$ dB, upper SNR bound $\gamma_{\max} = 30$ dB.

Output: Optimized trainable parameters Θ_{train}^* .

- 1: Initialize Θ_{train} using Θ_0 .
 - 2: Stage 1: basic detection fine-tuning without channel noise.
 - 3: **for** $r = 1$ to E_1 **do**
 - 4: Sample a mini-batch $(\mathbf{X}, Y_{\text{gt}})$ from $\mathcal{D}$.
 - 5: Generate initial detections $\mathcal{B}_{\text{init}}$.
 - 6: Compute the detection loss $\mathcal{L}_{\text{task}}(\mathcal{B}_{\text{init}}, Y_{\text{gt}})$.
 - 7: Update Θ_{train} by back-propagation.
 - 8: **end for**
 - 9: Load the best checkpoint obtained in Stage 1.
 - 10: Stage 2: noise-aware curriculum fine-tuning.
 - 11: **for** $e = 1$ to E_2 **do**
 - 12: $\gamma_{\min}(e) = \gamma_{\text{start}} - \frac{e-1}{E_2-1}(\gamma_{\text{start}} - \gamma_{\text{end}})$.
 - 13: Sample $\gamma \sim \mathcal{U}(\gamma_{\min}(e), \gamma_{\max})$.
 - 14: Sample a mini-batch $(\mathbf{X}, Y_{\text{gt}})$ from $\mathcal{D}$.
 - 15: Extract multi-scale features $\mathcal{F} = \{\mathbf{F}^{(l)}\}_{l=1}^L$ and time-scale parameter maps.
 - 16: Generate the semantic mask $\{\mathbf{M}^{(l)}\}_{l=1}^L$ using the threshold τ_m .
 - 17: Obtain sparse semantic features $\Psi = \{\mathbf{F}^{(l)} \odot \mathbf{M}^{(l)}\}_{l=1}^L$.
 - 18: Pack Ψ into the channel input $\mathbf{S}$.
 - 19: Obtain the noisy representation $\hat{\mathbf{S}}$ under the sampled SNR γ .
 - 20: Decode $\hat{\mathbf{S}}$, reconstruct $\hat{\mathcal{F}}$, and generate initial detections $\mathcal{B}_{\text{init}}$.
 - 21: Compute the detection loss $\mathcal{L}_{\text{task}}(\mathcal{B}_{\text{init}}, Y_{\text{gt}})$.
 - 22: Update Θ_{train} by back-propagation.
 - 23: **end for**
 - 24: **return** Θ_{train}^* .
-

6. Experimental Evaluation and Discussion

6.1. Experimental Setup

This section evaluates Mamba-KGSC through simulation-based V2V cooperative object detection experiments. First, it introduces the experimental platforms, dataset construction, evaluation metrics, and baseline methods. Subsequently, it analyzes the relationship between semantic transmission overhead and detection accuracy under different compression ratios. Furthermore, it compares the detection robustness of different perception–communication paradigms under various SNR conditions and different channel models, including AWGN, Rayleigh fading, and Rician fading channels, at a unified compression ratio. Finally, it validates the effectiveness and lightweight advantages of the multi-source knowledge base constraint verification module through ablation studies and complexity analysis. Unless otherwise specified, all methods are evaluated using the same data splits, the same detection category mappings, and the same mAP@0.5:0.95 metric to ensure the fairness of the comparison results.

6.1.1. Hardware and Software Platform

The experiments are run on a workstation with an Intel Core i7 CPU and an NVIDIA GeForce RTX 3060 GPU, so all compared methods are tested under identical hardware settings. The autonomous-driving co-simulation environment is built based on OpenCDA, utilizing CARLA 0.9.12 as the platform for generating traffic scenes, vehicle motions, and sensor data. The neural network modules are implemented based on PyTorch 2.1.2, with CUDA 12.1 used for GPU acceleration. The main implementation settings of the proposed framework are summarized in Table 4. These settings are used to improve the reproducibility of the training, semantic compression, and receiver-side knowledge verification processes.

Table 4. Main implementation settings of Mamba-KGSC.

Item	Setting
Input image resolution	$640 \times 640 \times 3$
Basic fine-tuning epochs	180
Noise-aware curriculum fine-tuning epochs	120
Batch size	8
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Final learning-rate factor	0.01
Mixed-precision training	Enabled
SNR curriculum range	$\gamma_{\min} : 5 \text{ dB} \rightarrow -10 \text{ dB}, \gamma_{\max} = 30 \text{ dB}$
Compression ratio settings	10%, 30%, 50%, 70%, and 90%
Selected compression ratio	50%

6.1.2. Dataset and Task-Oriented Resampling

The foundational visual dataset used for training and evaluation in this paper is derived from the MS-COCO dataset [42]. This choice is also consistent with recent object-detection-oriented semantic communication studies that adopt MS-COCO or COCO-trained object detectors to evaluate task-relevant visual semantic transmission [35,36]. Traffic participant categories relevant to V2V cooperative object detection are selected, including pedestrians, vehicles, motorcycles, and bicycles. This category setting remains consistent with the cooperative perception task configuration in the OpenCDA co-simulation environment [43].

Directly using the original COCO traffic-related subset introduces a distinct category distribution imbalance. Specifically, the number of images containing only pedestrians

is significantly higher than that of complex traffic scene images containing vehicle-class targets, which easily causes the model to bias towards high-frequency pedestrian features during training while weakening the effective modeling of traffic targets such as vehicles, motorcycles, and bicycles. For the endogenous semantic encoding mechanism in this paper, this category bias may also affect the learning of spatial semantic masks, causing the model to lean more towards preserving high-frequency target regions while reducing the feature preservation capability for minority traffic categories.

To alleviate the aforementioned issues, this paper adopts a scene-aware category resampling strategy to filter out redundant pure-pedestrian images from the training subset and preserve samples containing vehicle-class targets as much as possible. The dataset distribution before and after resampling is shown in Table 5. This strategy reduces training sample redundancy while increasing the relative proportion of traffic target categories, helping the model learn more balanced task-relevant semantic representations.

Table 5. Dataset distribution before and after scene-aware category resampling.

Item	Original COCO Traffic Subset	Resampled Dataset	Reduction Ratio
Total Images	117,266	46,710	60.18%
Person	257,252	165,967	35.48%
Vehicle	64,693	61,547	4.86%
Motorcycle	8654	8654	0.00%
Bicycle	7056	7056	0.00%

As shown in Table 5, after resampling, the total number of images is reduced by 60.18%, the number of pedestrian instances is reduced by 35.48%, while vehicle instances are reduced by only 4.86%. The motorcycle and bicycle categories are not filtered, with a reduction ratio of 0.00%. These results indicate that the adopted resampling strategy can reduce training redundancy while preserving traffic target samples closely related to V2V cooperative object detection as much as possible, providing a relatively balanced data foundation for subsequent compression robustness and detection accuracy evaluations.

6.1.3. Evaluation Metrics

To evaluate detection accuracy, communication efficiency, and physical consistency in V2V cooperative object detection, four groups of metrics are used.

(1) Mean Average Precision (mAP@0.5:0.95). The IoU thresholds used for this metric are collected in $\mathcal{T}$:

$$\mathcal{T} = \{0.50, 0.55, \dots, 0.95\}. \quad (36)$$

Let AP_{τ} be the average precision under the IoU threshold $\tau \in \mathcal{T}$. The mAP@0.5:0.95 metric is computed as:

$$mAP@0.5 : 0.95 = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} AP_{\tau}. \quad (37)$$

Compared to the evaluation method using only mAP@0.5, mAP@0.5:0.95 imposes higher requirements on bounding box localization accuracy and can more strictly reflect the spatial localization capability of the model in autonomous driving object detection tasks.

(2) Physical Consistency Filtering Statistics. To characterize the suppression effect of the multi-source knowledge base on abnormal detection results and repeated detection results, this paper counts the numerical changes in the initial detection set after physical consistency verification and visual–physical spatial joint redundancy suppression. Specifically, let the initial detection set be $\mathcal{B}_{init}$, the candidate set after physical consistency constraint screening be $\mathcal{B}_{valid}$, and the final output set after visual–physical spatial

joint non-maximum suppression be $\mathcal{B}_{\text{final}}$. The remaining number of detections, the number of filtered or merged candidate bounding boxes at each stage, and the reduction or improvement ratios relative to the baseline method can be counted respectively. These statistical results are used to reflect the effectiveness of the multi-source knowledge base in eliminating physically inconsistent abnormal candidate bounding boxes and suppressing repeated detections.

To measure the stability of the number of detection outputs under extremely low signal-to-noise ratios (SNRs), this paper further calculates the variance of the number of detection bounding boxes. Let the test sequence contain a total of N_f frames, and the number of output detection bounding boxes in the t -th frame be n_t . The average number of detection bounding boxes per frame is:

$$\bar{n} = \frac{1}{N_f} \sum_{t=1}^{N_f} n_t. \quad (38)$$

Then the variance of the number of detection bounding boxes is defined as:

$$\text{Var}_{\text{det}} = \frac{1}{N_f} \sum_{t=1}^{N_f} (n_t - \bar{n})^2, \quad (39)$$

where a larger Var_{det} indicates more obvious fluctuations in the number of detection bounding boxes across different frames. A smaller value indicates more stable detection output quantities.

(3) Transmission Bandwidth Ratio (TBR). Let the spatial semantic mask at the l -th scale be $\mathbf{M}^{(l)} \in \{0, 1\}^{H_l \times W_l}$; then, its number of valid preserved locations can be represented by $\|\mathbf{M}^{(l)}\|_0$. The overall transmission bandwidth ratio is defined as:

$$\text{TBR} = \frac{\sum_{l=1}^L \|\mathbf{M}^{(l)}\|_0}{\sum_{l=1}^L H_l W_l} \times 100\%, \quad (40)$$

where H_l and W_l denote the height and width of the feature map at the l -th scale, respectively, and L denotes the number of multi-scale feature layers. Taking the semantic feature vector at a spatial location as the statistical unit, this metric reflects the proportion of non-zero semantic feature vectors ultimately transmitted by the transmitter relative to the original dense feature locations. A lower value indicates stronger compression. The corresponding Compression Ratio (CR) is defined as:

$$\text{CR} = 100\% - \text{TBR}. \quad (41)$$

For ease of presentation, subsequent experiments mainly use the compression ratio CR as the control variable, which has a one-to-one correspondence with the transmission bandwidth ratio TBR.

(4) Computational Complexity. This paper evaluates the model size and computational overhead of different methods from two perspectives: total model parameters and total floating-point operations. Specifically, the total number of model parameters reflects the scale of the trainable modules in the perception–communication pipeline, while FLOPs indicate the computational cost of a single forward inference. Let Θ be the set of trainable parameters in the module to be counted. Its parameter quantity is defined as:

$$\text{Params} = \sum_{\theta \in \Theta} \text{numel}(\theta), \quad (42)$$

where θ represents a trainable parameter tensor in the module, and $\text{numel}(\theta)$ represents the number of scalar parameters contained in this tensor. The FLOPs for a single forward inference are obtained using standard complexity-statistics tools.

The above metrics are used to measure the differences between different perception–communication paradigms in terms of model size and computational overhead.

6.1.4. Baseline Methods

To verify the effectiveness of the proposed Mamba-KGSC framework in the V2V cooperative object detection task, this paper compares it with two representative baseline methods under the same dataset splits, the same detection category mappings, the same SNR settings, and the same mAP@0.5:0.95 metric.

1. Separated SSCC baseline method.

This baseline adopts the traditional separated source–channel coding architecture, consisting of JPEG source coding, LDPC channel coding, wireless channel transmission, channel decoding, image reconstruction, and the YOLOv5 detector. The transmitter first uses JPEG to compress the input image and calculates the number of bits required for the compressed bitstream. Subsequently, it employs LDPC as a representative channel coding scheme to protect the compressed bitstream. This baseline is primarily used to characterize the cliff effect that may occur in traditional SSCC links under low-SNR or resource-constrained conditions. For the fading-channel evaluation, the separated communication link is further tested under Rayleigh and Rician channel conditions to analyze its robustness against typical flat-fading perturbations.

2. DeepJSCC baseline method.

This baseline follows the DeepJSCC idea of jointly learning source representation and channel mapping, and it is adapted here to the object detection setting [44]. Specifically, the transmitter first extracts the multi-scale visual features of the YOLOv5 backbone network and compresses them into a continuous channel input representation via a neural encoder. Subsequently, additive Gaussian noise corresponding to the SNR is superimposed on this representation to simulate wireless channel perturbations. In the fading-channel evaluation, Rayleigh or Rician multiplicative fading is further applied before additive noise injection, so as to examine the robustness of the learned feature transmission mechanism under typical flat-fading conditions. The receiver recovers the multi-scale features through the corresponding neural decoder and inputs them into the YOLOv5 detection head to complete object detection. To ensure comparability at the communication overhead level, this paper evaluates the detection performance of DeepJSCC under different channel bandwidth ratio settings, and selects the 50% compression ratio as the primary comparison condition in subsequent comparative experiments.

6.2. Detection Accuracy Under Different Compression Ratios

To analyze the impact of the compression ratio setting on detection performance and provide a unified communication overhead benchmark for subsequent comparative experiments under different SNR conditions and channel conditions, this paper evaluates the mAP@0.5:0.95 variations in Mamba-KGSC and the DeepJSCC baseline method under different compression ratios, respectively. Because the compression control mechanisms of the two methods are different, this paper does not conduct a point-by-point equivalent comparison, but instead separately analyzes their performance variation trends under their respective compression mechanisms. Specifically, Mamba-KGSC controls the proportion of valid transmission features by adjusting the spatial semantic mask threshold τ_m , and thus five sets of compression ratios are set: 10%, 30%, 50%, 70%, and 90%. This experiment can

also be regarded as a sensitivity analysis of the semantic mask threshold, since different values of τ_m lead to different spatial mask sparsity levels and therefore different transmission bandwidth ratios. The DeepJSCC baseline method controls the channel bandwidth ratio by adjusting the bottleneck bandwidth of the feature autoencoder, and thus three sets of typical compression ratios are set: 12.5%, 25%, and 50%.

It should be noted that Figure 4 focuses on compression-ratio sensitivity analysis under a unified AWGN channel setting.

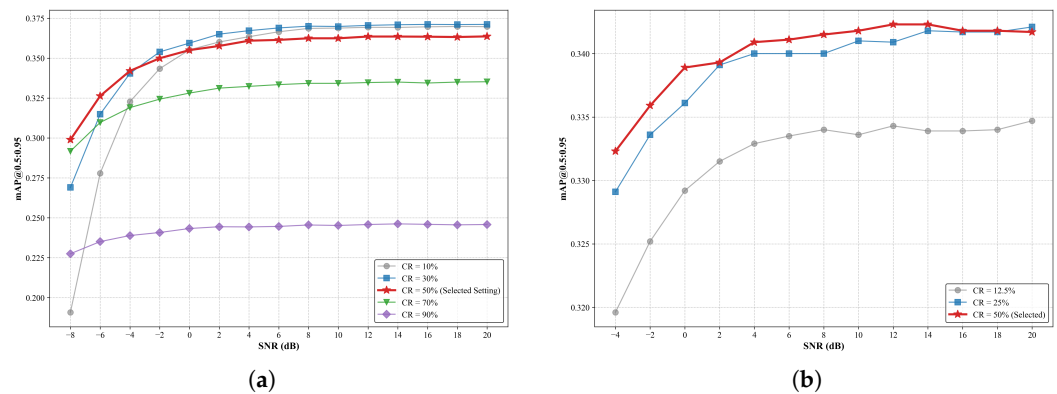


Figure 4. Detection accuracy variations under different compression ratio settings. (a) Mamba-KGSC under different semantic compression ratios. (b) DeepJSCC baseline under different compression ratios.

In Figure 4a, the detection accuracy under the 10%, 30%, and 50% compression ratios generally remains at a high level, while the mAP@0.5:0.95 under the 70% and 90% compression ratios drops significantly, indicating that excessively strong sparse encoding weakens the foreground object boundaries and local structure information. In contrast, the 50% compression ratio maintains relatively stable detection performance in both the low-SNR and the medium-to-high SNR intervals, indicating that this setting can achieve a good empirical balance between reducing transmission overhead and preserving task-relevant semantic features.

Figure 4b reports how the DeepJSCC baseline changes as the compression ratio varies. Under stronger compression, DeepJSCC shows a more pronounced accuracy drop in the low-SNR interval. This indicates that under strong compression conditions, the external neural encoder may struggle to adequately preserve fine-grained information related to object boundaries and spatial localization, thereby limiting the bounding box regression accuracy of the detection head at the receiver. As the compression ratio decreases, the detection performance of DeepJSCC gradually improves, but its independent neural codec structure still introduces a higher model size and computational overhead.

Synthesizing the results of Figure 4a,b, this paper selects the 50% compression ratio as the unified setting for subsequent comparative experiments. This setting does not represent the globally optimal compression ratio, but rather exhibits a superior trade-off between transmission overhead and detection accuracy within the experimental scope of this paper. Therefore, the 50% compression ratio is selected as an empirical operating point that avoids excessive semantic feature removal while still reducing the transmission burden. This setting can provide a consistent communication overhead benchmark for fair comparisons under different SNR conditions and channel models in the subsequent sections.

6.3. Detection Accuracy Under Different SNRs and Channel Conditions

Under the unified 50% compression ratio setting, this paper compares the detection accuracy variations in the traditional SSCC, the DeepJSCC baseline method, and the front-end semantic communication branch of Mamba-KGSC under different SNR conditions across

AWGN, Rayleigh fading, and Rician fading channels. All methods adopt mAP@0.5:0.95 as the evaluation metric to measure the comprehensive performance of object boundary localization and category detection under varying noise and fading perturbations. For the Rician fading channel, the Rician K -factor is set to 2 on a linear scale.

It should be noted that the COCO dataset does not contain map semantics, camera poses, and semantic raycasting information aligned with OpenCDA/CARLA. Therefore, the mAP evaluation in this section primarily targets the front-end semantic encoding and the initial detection branch at the receiver of Mamba-KGSC. The role of the multi-source knowledge base constraint verification module will be evaluated separately in the subsequent ablation studies and visualization analysis. Thus, the results in this subsection mainly reflect the robustness of the front-end semantic communication branch under different channel conditions.

As can be seen from Figure 5a–c, the traditional SSCC baseline method exhibits a sharp performance drop under low-SNR conditions across all three channel settings. When the SNR is below 0 dB, the source recovery quality in the separated communication link drops significantly, leading to an obvious degradation in the quality of the reconstructed image at the receiver, which in turn causes a rapid decline in the localization performance of the YOLOv5 detector. When the SNR drops to -4 dB, the mAP@0.5:0.95 of this method has fallen to a relatively low level, reflecting the typical cliff effect of traditional separated communications under harsh channels. This phenomenon is particularly evident under fading channels, where multiplicative channel fluctuations further affect the reliability of separated source–channel transmission.

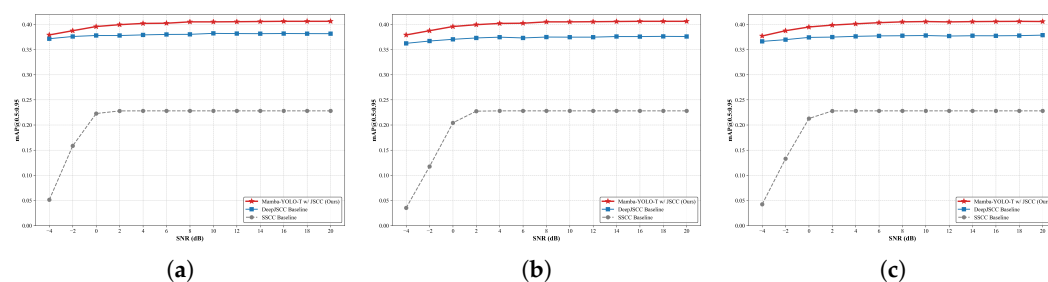


Figure 5. SNR–mAP performance comparison of different perception–communication paradigms under the 50% compression ratio: (a) AWGN channel; (b) Rayleigh fading channel; (c) Rician fading channel.

The DeepJSCC baseline method exhibits smoother performance variations under low-SNR conditions, indicating that the end-to-end learned joint source–channel coding can alleviate the sudden performance degradation of traditional SSCC to a certain extent. Compared with AWGN, its performance under Rayleigh and Rician fading channels shows a slight decrease, but the overall curve remains continuous without an obvious cliff effect. However, under the strict mAP@0.5:0.95 metric, its detection accuracy remains limited under all three channel conditions. This is likely because DeepJSCC relies on an external autoencoder to compress and recover the overall features. At high compression ratios, it is prone to weakening fine-grained information related to object boundaries, local textures, and spatial locations, thereby limiting the high-precision bounding box regression capability of the detection head.

By contrast, the front-end semantic communication branch of Mamba-KGSC keeps its detection performance relatively stable over the tested SNR range under AWGN, Rayleigh, and Rician channel conditions. In the low-SNR region in particular, the proposed method still achieves a higher mAP@0.5:0.95 than both baseline methods, and the performance gap remains observable under fading channels. This is mainly attributed to the endogenous semantic sparse encoding mechanism at the transmitter: the system does not compress all

features indiscriminately, but instead suppresses background redundant features through a semantic preservation probability-driven mask gating strategy, allowing limited channel resources to be prioritized for transmitting semantic features related to foreground objects. Therefore, even when fading-induced feature perturbations are introduced, the transmitted sparse semantic stream still retains more detection-related foreground, boundary, and localization cues. It should also be noted that the multi-source knowledge base module at the receiver is mainly used for physical consistency verification and visual–physical spatial joint redundancy suppression in the subsequent OpenCDA/CARLA scenes, and its effect will be further analyzed in Section 6.4.

In summary, under the same 50% compression ratio setting, Mamba-KGSC exhibits more stable detection accuracy compared with the traditional SSCC and DeepJSCC baseline methods across AWGN, Rayleigh, and Rician channel conditions. These results mainly demonstrate the robustness of the front-end endogenous semantic encoding branch under noisy and fading channels. The contribution of the receiver-side multi-source knowledge base constraint verification module to abnormal detection suppression and repeated-box reduction will be further analyzed in the following ablation studies and visualization experiments.

6.4. Ablation Analysis of Multi-Source Knowledge Base Constraint Verification

To evaluate the noise resistance capability and result verification effect of Mamba-KGSC under harsh communication conditions, this section conducts ablation experiments under the -10 dB low-SNR condition. The experiments separately analyze the statistical impact of the front-end noise-aware semantic training on the stability of the detection output, as well as the suppression effect of the multi-source knowledge base constraint verification at the receiver on abnormal candidate bounding boxes and repeated candidate bounding boxes. These ablation results correspond to the two main components of the proposed framework, namely the endogenous semantic encoding branch at the transmitter and the multi-source knowledge base constraint verification module at the receiver, thereby clarifying their individual contributions to noise robustness, abnormal bounding-box suppression, and detection-result refinement under the same communication-overhead setting.

As shown in Table 6, under the -10 dB low-SNR condition, the number of detection bounding boxes output by the Mamba-YOLO-T front-end without JSCC training increases significantly, with the total number of detections reaching 56,944 and the average number of detections per frame being 4.75. After introducing the front-end JSCC training, the total number of detections drops to 17,807, and the average number of detections per frame drops to 1.48, corresponding to a 68.7% reduction in the total number of candidate detections. Since count variance is scale-dependent, it is used here only as an auxiliary indicator of absolute frame-level fluctuation. The absolute count variance decreases from 3.56 to 2.07, indicating that the frame-level fluctuation of candidate detections is also reduced after JSCC training. Overall, this result suggests that task-oriented JSCC training can reduce noise-induced redundant candidate detections and provide a more compact and reliable initial candidate set for receiver-side verification.

It should be pointed out that the front-end JSCC training primarily enhances the noise resistance capability of the features, but it cannot fully guarantee that the detection results conform to the physical constraints in traffic scenes. The front-end JSCC ablation in Table 6 and the receiver-side knowledge-base ablation in Table 7 are conducted over different evaluation durations and are designed to examine different functional components. Therefore, the absolute candidate counts in the two tables are not intended for direct one-to-one comparison. In Table 7, the 13,728 candidate bounding boxes are used as the front-end-only baseline for evaluating the subsequent receiver-side knowledge constraint

verification process. These candidate bounding boxes may still contain abnormal results with unreasonable physical locations and repeated detections caused by fragmented local feature responses. To further examine the effect of the multi-source knowledge base module, the receiver-side constraint verification process is evaluated in two stages: physical consistency screening based on semantic anchor mapping, followed by visual–physical spatial joint non-maximum suppression. The corresponding statistical results are shown in Table 7.

Table 6. Ablation statistical results of the front-end endogenous semantic encoding branch with and without JSCC training under the -10 dB low-SNR condition.

Model Paradigm	Total Detections	Avg./Frame	Count Variance	Candidate Count Reduction
Mamba-YOLO-T (without JSCC)	56,944	4.75	3.56	Baseline
Mamba-YOLO-T (with JSCC)	17,807	1.48	2.07	68.7%

Table 7. Ablation statistical results of the multi-source knowledge base constraint verification module by stage.

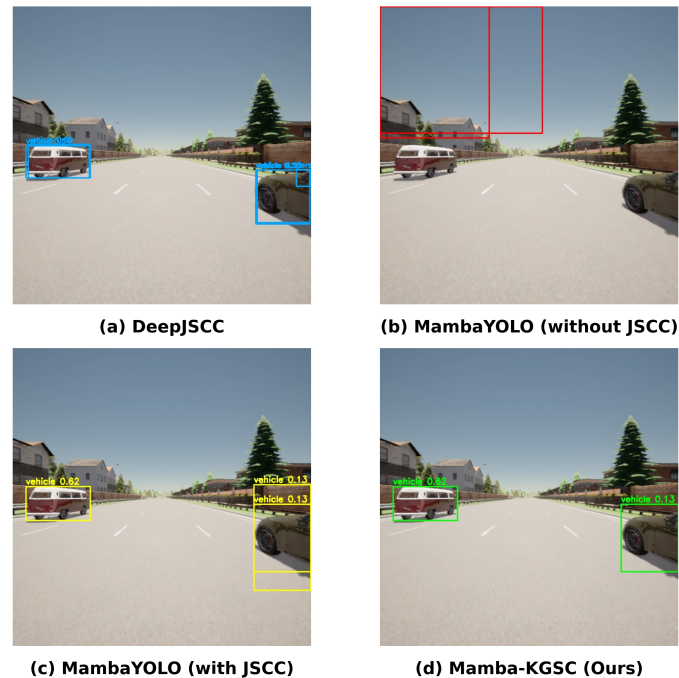
Enabled Screening Stage	Remaining Detections	Filtered/Merged Boxes	Cumulative Reduction Ratio
Mamba-KGSC front-end only	13,728	–	Baseline
Physical Consistency Screening	11,434	2294	16.7% Reduction
Visual–Physical Joint NMS	9370	2064	31.7% Reduction

As shown in Table 7, the physical consistency screening stage reduces the number of candidate bounding boxes from 13,728 to 11,434, filtering out a total of 2294 physically inconsistent candidate bounding boxes, with a reduction ratio of 16.7%. This stage mainly relies on the drivable-area semantic constraint, the 2D geometric envelope constraint, and the semantic raycasting constraint, which are used to eliminate candidate targets that fall into non-navigable regions, exhibit abnormal geometric shapes, or are inconsistent with map semantics. Subsequently, the visual–physical spatial joint non-maximum suppression further reduces the number of candidate bounding boxes to 9370, merging a total of 2064 repeated candidate bounding boxes, bringing the total reduction ratio to 31.7%. These results indicate that the multi-source knowledge base module provides an additional and independent refinement effect after the front-end semantic communication branch, especially in suppressing physically inconsistent candidate boxes and repeated detections at the receiver.

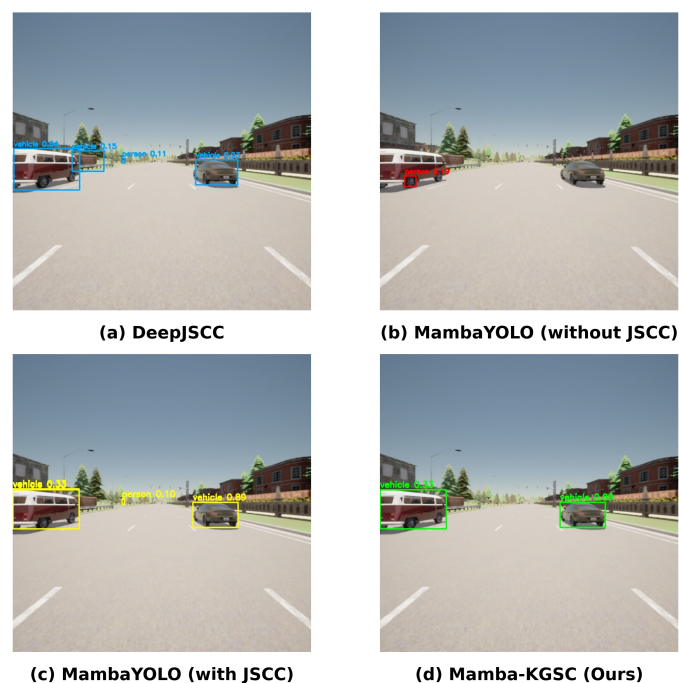
To intuitively demonstrate the detection differences among different perception–communication paradigms under extreme channel conditions, this paper selects two typical scenes for visualization comparison, as shown in Figure 6.

Figure 6A illustrates a repeated detection scene caused by fragmented local feature responses. Although DeepJSCC and the JSCC-trained Mamba-YOLO-T front-end can detect the targets, they may still generate multiple candidate bounding boxes near the same target. In contrast, Mamba-KGSC effectively reduces repeated detections through visual–physical spatial joint non-maximum suppression, which jointly utilizes the 2D overlap relationship and the 3D physical anchor distance for candidate bounding box merging.

Figure 6B displays an abnormal detection scene under severe noise perturbations. The front-end model without noise-aware training is prone to misidentifying noise-perturbed regions as targets that do not exist in the physical space. Although the JSCC-trained front-end improves overall stability, it may still produce candidate bounding boxes with unstable boundaries or unreasonable physical locations.



(A) Repeated detection scene.



(B) Abnormal detection scene.

Figure 6. Visualization analysis of detection results under extreme channel conditions in OpenCDA scenes.

Mamba-KGSC further utilizes the multi-source knowledge base module to perform physical consistency verification on candidate targets, which can eliminate abnormal detection results that clearly violate road semantics and scene constraints, outputting more stable detection results.

6.5. Computational Complexity

To evaluate the model size and computational overhead of different perception–communication paradigms, this paper calculates the total number of model parameters and the FLOPs for a single forward pass in the inference pipeline of each method, with the results shown in Table 8. All statistics are obtained under the same input scale and detection category setting, where the multi-source knowledge base constraint verification module is a rule-based result verification process and is not included in the neural network parameter statistics or neural FLOP calculation.

The traditional SSCC baseline method adopts YOLOv5 as the detector, with 7.23 M parameters and 8.22 GFLOPs for a single forward pass. This baseline has a lower computational overhead, but its performance stability under low-SNR conditions shows obvious limitations in the aforementioned SNR experiments.

The DeepJSCC baseline method also adopts the YOLOv5 detector and completes the joint source–channel mapping through an independent neural codec structure. This structure increases the model size to 23.45 M parameters and the computational cost to 22.75 GFLOPs for a single forward pass, indicating that while it improves channel robustness, it also incurs a larger model size and higher computational overhead.

In contrast, Mamba-KGSC adopts Mamba-YOLO-T as the perception backbone at the transmitter and uses the internal time-scale responses of the backbone network to generate spatial semantic masks, thereby realizing task-relevant semantic feature screening and sparse semantic encoding. The model contains 5.99 M parameters and requires 13.60 GFLOPs for a single forward pass. Its FLOPs are higher than those of the traditional SSCC baseline, but remain substantially below those of the DeepJSCC baseline. This result shows that Mamba-KGSC provides a favorable trade-off among model size, computational overhead, transmission efficiency, and detection stability.

Table 8. Comparison of computational complexity among different methods.

Method	Total Model Params (M)	Total FLOPs (G)
Traditional SSCC	7.23	8.22
DeepJSCC baseline	23.45	22.75
Mamba-KGSC (Ours)	5.99	13.60

7. Conclusions

Targeting the V2V cooperative object detection task under limited transmission resources and noisy channel conditions, this paper proposes a knowledge-guided lightweight semantic communication framework, named Mamba-KGSC. This framework unifies task-relevant semantic encoding at the transmitter, wireless channel transmission, semantic decoding at the receiver, and multi-source knowledge base constraint verification into the same cooperative perception pipeline. The transmitter utilizes the internal time-scale parameters of the Mamba-YOLO-T backbone network to generate spatial semantic masks, realizing the selective preservation and sparse semantic encoding of foreground object-related features. The receiver integrates static map knowledge, geometric priors, spatial topological constraints, and visual–physical spatial association rules to perform physical consistency screening and joint redundancy suppression on the initial detection results, thereby enhancing the reliability of the detection results under harsh channels.

Experimental results demonstrate that Mamba-KGSC exhibits relatively stable detection performance under varying compression ratios, SNR conditions, and channel models. Under a 50% compression ratio and low-SNR conditions, the proposed method achieves higher mAP@0.5:0.95 detection accuracy compared with the traditional SSCC and DeepJSCC baseline methods. Meanwhile, the multi-source knowledge base constraint verification can effectively reduce physically inconsistent abnormal detections and repeated detections. Computational complexity analysis further indicates that Mamba-KGSC achieves a favorable trade-off among transmission overhead, detection accuracy, model size, and physical consistency, providing a potential implementation scheme for robust V2V cooperative object detection in harsh communication environments.

Nevertheless, this study still has several limitations. First, the proposed framework is mainly validated in an OpenCDA/CARLA-based software simulation environment, while experiments on real vehicular edge hardware have not yet been conducted. In addition, the SS2D-based Mamba backbone involves multi-directional scanning operations, and the current prototype has not incorporated dedicated CUDA or kernel-level acceleration. Therefore, the practical processing latency and deployment efficiency on resource-constrained edge platforms require further evaluation. Second, the receiver-side multi-source knowledge base verification module relies on HD maps, semantic raycasting, and vehicle pose information, which may be obtained from HD map services, cooperative localization, GNSS/INS, SLAM, or online map construction systems in practical V2V deployments. However, missing or outdated map semantics, localization noise, and calibration drift may affect physical-anchor estimation and scene-consistency verification. Third, although Rayleigh and Rician fading channels are considered, the current channel simulation mainly controls link quality through predefined SNR settings and does not explicitly couple distance-dependent path loss with the inter-vehicle distance in CARLA.

Future work will focus on edge-oriented acceleration and CUDA/kernel-level optimization. In addition, map uncertainty, pose perturbation, calibration drift, packet loss, transmission latency, Doppler effects, and distance-dependent path-loss models will be further incorporated into the evaluation framework. Hardware-in-the-loop and real-world V2V cooperative perception experiments will also be conducted to further validate the practical feasibility of Mamba-KGSC.

Author Contributions: Conceptualization, G.W.; methodology, G.W.; software, G.W.; validation, G.W., J.S., Y.L. and M.H.; formal analysis, G.W.; investigation, G.W. and J.S.; data curation, G.W. and Y.L.; writing—original draft preparation, G.W.; writing—review and editing, G.W., J.S., Y.L., M.H. and P.Z.; visualization, G.W.; experiments, G.W.; resources, J.S., M.H. and P.Z.; supervision, P.Z.; project administration, P.Z.; funding acquisition, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant U24A20211 and Grant 62271096; the Natural Science Foundation of Chongqing under Grant CSTB2025NSCQ-GPX1275; the Chongqing Municipal Education Commission under Grant KJQN202400628; in part by the China Postdoctoral Science Foundation under Grant 2022MD723723; and in part by the Chongqing Postdoctoral Research Program Special Grant under Grant 2023CQBSHTB3092.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors appreciate the constructive comments and suggestions provided by the editors and reviewers.

Conflicts of Interest: Apart from the disclosed editorial relationship, the authors declare no commercial or financial conflicts of interest. Prof. Dr. Puning Zhang is listed as a Guest Editor of this Special

Issue. Prof. Dr. Puning Zhang was not involved in the editorial handling, peer-review process, or decision-making process for this manuscript. The authors declare no other conflicts of interest.

References

- Wang, T.-H.; Manivasagam, S.; Liang, M.; Yang, B.; Zeng, W.; Urtasun, R. V2VNet: Vehicle-to-Vehicle Communication for Joint Perception and Prediction. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 605–621. https://doi.org/10.1007/978-3-030-58536-5_36.
- Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; Ma, J. OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 2583–2589. <https://doi.org/10.1109/ICRA46639.2022.9812038>.
- Liu, Y.-C.; Tian, J.; Glaser, N.; Kira, Z. When2com: Multi-Agent Perception via Communication Graph Grouping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4106–4115. <https://doi.org/10.1109/CVPR42600.2020.00416>.
- Hu, Y.; Fang, S.; Lei, Z.; Zhong, Y.; Chen, S. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 4874–4886.
- Fan, W.; Wei, N.; Bazzi, A.; Xi, R.; Song, Z.; Li, Y.; Zeng, Z.; Xiu, Y.; Assi, C. Optimizing Tracking Accuracy in Energy-Constrained Multimodal ISAC via Lyapunov-Driven Heterogeneous Mixture-of-Experts. *arXiv* **2026**, arXiv:2604.26330. <https://doi.org/10.48550/arXiv.2604.26330>.
- Zhang, P.; Hu, M.; Zhang, H.; Wu, C.; Yang, Z. Personality-Aware Multimodal Driver Emotion Recognition Towards Intelligent Connected Vehicles. *IEEE Trans. Affect. Comput.* **2026**, *17*, 801–816. <https://doi.org/10.1109/TAFFC.2025.3634565>.
- Zhang, P.; Wang, Y.; Wang, J.; Zhang, Z. Multi-Modal Encrypted Retrieval Method with Semantic Feature Fusion towards Internet of Medical Things. *IEEE J. Biomed. Health Inform.* **2025**, *30*, 4772–4785. <https://doi.org/10.1109/JBHI.2025.3638367>.
- Zhang, P.; Fu, M.; Zhao, R.; Zhang, H.; Luo, C. PURE: Personality-Coupled Multi-Task Learning Framework for Aspect-Based Multimodal Sentiment Analysis. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 462–477. <https://doi.org/10.1109/TKDE.2024.3485108>.
- Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
- Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
- Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.-H. Deep Learning Enabled Semantic Communication Systems. *IEEE Trans. Signal Process.* **2021**, *69*, 2663–2675. <https://doi.org/10.1109/TSP.2021.3071210>.
- Bourtsoulatz, E.; Kurka, D.B.; Gündüz, D. Deep Joint Source–Channel Coding for Wireless Image Transmission. *IEEE Trans. Cogn. Commun. Netw.* **2019**, *5*, 567–579. <https://doi.org/10.1109/TCCN.2019.2919300>.
- Kurka, D.B.; Gündüz, D. DeepJSCC-f: Deep Joint Source–Channel Coding of Images with Feedback. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 178–193. <https://doi.org/10.1109/JSAIT.2020.2987203>.
- Kurka, D.B.; Gündüz, D. Bandwidth-Agile Image Transmission with Deep Joint Source–Channel Coding. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 8081–8095. <https://doi.org/10.1109/TWC.2021.3090048>.
- Lyu, Z.; Zhu, G.; Xu, J.; Ai, B.; Cui, S. Semantic Communications for Image Recovery and Classification via Deep Joint Source and Channel Coding. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 8388–8404. <https://doi.org/10.1109/TWC.2023.3349330>.
- Gu, A.; Goel, K.; Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
- Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In Proceedings of the First Conference on Language Modeling (COLM), Philadelphia, PA, USA, 7–9 October 2024.
- Wang, Z.; Li, C.; Xu, H.; Zhu, X.; Li, H. Mamba YOLO: A Simple Baseline for Object Detection with State Space Model. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 8205–8213. <https://doi.org/10.1609/aaai.v39i8.32885>.
- Yang, B.; Liang, M.; Urtasun, R. HDNET: Exploiting HD Maps for 3D Object Detection. In Proceedings of the 2nd Conference on Robot Learning (CoRL), Zürich, Switzerland, 29–31 October 2018; pp. 146–155.
- Li, Q.; Wang, Y.; Wang, Y.; Zhao, H. HDMapNet: An Online HD Map Construction and Evaluation Framework. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 4628–4634. <https://doi.org/10.1109/ICRA46639.2022.9812383>.
- Liao, B.; Chen, S.; Wang, X.; Cheng, T.; Zhang, Q.; Liu, W.; Huang, C. MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
- Chen, Q.; Tang, S.; Yang, Q.; Fu, S. Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds. In Proceedings of the IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019; pp. 514–524. <https://doi.org/10.1109/ICDCS.2019.00058>.

23. Chen, Q.; Ma, X.; Tang, S.; Guo, J.; Yang, Q.; Fu, S. F-Cooper: Feature-Based Cooperative Perception for Autonomous Vehicle Edge Computing System Using 3D Point Clouds. In Proceedings of the ACM/IEEE Symposium on Edge Computing (SEC), Arlington, VA, USA, 7–9 November 2019; pp. 88–100. <https://doi.org/10.1145/3318216.3363300>.
24. Li, Y.; Ma, D.; An, Z.; Wang, Z.; Zhong, Y.; Chen, S.; Feng, C. V2X-Sim: Multi-Agent Collaborative Perception Dataset and Benchmark for Autonomous Driving. *IEEE Robot. Autom. Lett.* **2022**, *7*, 10914–10921. <https://doi.org/10.1109/LRA.2022.3192802>.
25. Xu, R.; Xiang, H.; Tu, Z.; Xia, X.; Yang, M.-H.; Ma, J. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 107–124. https://doi.org/10.1007/978-3-031-19842-7_7.
26. Xu, R.; Tu, Z.; Xiang, H.; Shao, W.; Zhou, B.; Ma, J. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. In Proceedings of the Conference on Robot Learning (CoRL), Auckland, New Zealand, 14–18 December 2022; pp. 989–1000.
27. Lu, Y.; Li, Q.; Liu, B.; Dianati, M.; Feng, C.; Chen, S.; Wang, Y. Robust Collaborative 3D Object Detection in Presence of Pose Errors. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 4812–4818. <https://doi.org/10.1109/ICRA48891.2023.10160546>.
28. Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; Zhang, L. How2comm: Communication-Efficient and Collaboration-Pragmatic Multi-Agent Perception. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 25151–25164.
29. Li, Y.; Ren, S.; Wu, P.; Chen, S.; Feng, C.; Zhang, W. Learning Distilled Collaboration Graph for Multi-Agent Perception. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 29541–29552.
30. Yang, K.; Yang, D.; Zhang, J.; Wang, H.; Sun, P.; Song, L. What2comm: Towards Communication-Efficient Collaborative Perception via Feature Decoupling. In Proceedings of the 31st ACM International Conference on Multimedia (ACM MM), Ottawa, ON, Canada, 29 October–3 November 2023; pp. 7686–7695. <https://doi.org/10.1145/3581783.3611699>.
31. Huang, T.; Liu, J.; Zhou, X.; Nguyen, D.C.; Azghadi, M.R.; Xia, Y.; Han, Q.-L.; Sun, S. Vehicle-to-Everything Cooperative Perception for Autonomous Driving. *Proc. IEEE* **2025**, *113*, 443–477. <https://doi.org/10.1109/JPROC.2025.3600903>.
32. Gündüz, D.; Qin, Z.; Aguerri, I.E.; Dhillon, H.S.; Yang, Z.; Yener, A.; Wong, K.K.; Chae, C.-B. Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications. *IEEE J. Sel. Areas Commun.* **2023**, *41*, 5–41. <https://doi.org/10.1109/JSAC.2022.3223408>.
33. Lu, Z.; Li, R.; Lu, K.; Chen, X.; Hossain, E.; Zhao, Z.; Zhang, H. Semantics-Empowered Communications: A Tutorial-Cum-Survey. *IEEE Commun. Surv. Tutor.* **2024**, *26*, 41–79. <https://doi.org/10.1109/COMST.2023.3333342>.
34. Yang, W.; Du, H.; Liew, Z.Q.; Lim, W.Y.B.; Xiong, Z.; Niyato, D.; Chi, X.; Shen, X.; Miao, C. Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 213–250. <https://doi.org/10.1109/COMST.2022.3223224>.
35. Guo, J.; Yin, H.; Song, B.; Chi, Y.; Zhang, Z.; Yuen, C.; Niyato, D. Multi-Scale Semantic Communication for Object Detection: Single and Cross-Domain Scenarios. *IEEE Trans. Wirel. Commun.* **2025**, *24*, 6195–6210. <https://doi.org/10.1109/TWC.2025.3552255>.
36. Safaeipour, F.Z.; Hashemi, M. Semantic-Aware and Goal-Oriented Communications for Object Detection in Wireless End-to-End Image Transmission. In Proceedings of the 2024 International Conference on Computing, Networking and Communications (ICNC), Big Island, HI, USA, 19–22 February 2024; pp. 182–187. <https://doi.org/10.1109/ICNC59896.2024.10556058>.
37. Tung, T.-Y.; Kurka, D.B.; Jankowski, M.; Gündüz, D. DeepJSCC-Q: Constellation Constrained Deep Joint Source–Channel Coding. *IEEE J. Sel. Areas Inf. Theory* **2022**, *3*, 720–731. <https://doi.org/10.1109/JSAIT.2022.3231042>.
38. Shao, J.; Mao, Y.; Zhang, J. Task-Oriented Communication for Multidevice Cooperative Edge Inference. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 73–87. <https://doi.org/10.1109/TWC.2022.3191118>.
39. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
40. Pan, Y.A.; Li, F.; Li, A.; Niu, Z.; Liu, Z. Urban Intersection Traffic Flow Prediction: A Physics-Guided Stepwise Framework Utilizing Spatio-Temporal Graph Neural Network Algorithms. *Multimodal Transp.* **2025**, *4*, 100207. <https://doi.org/10.1016/j.multra.2025.100207>.
41. Yang, J.; Lu, J.; Lee, S.; Batra, D.; Parikh, D. Graph R-CNN for Scene Graph Generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–685.
42. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zürich, Switzerland, 6–12 September 2014; pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

43. Xu, R.; Guo, Y.; Han, X.; Xia, X.; Xiang, H.; Ma, J. OpenCDA: An Open Cooperative Driving Automation Framework Integrated with Co-Simulation. In Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 1155–1162. <https://doi.org/10.1109/ITSC48978.2021.9564825>.
44. Zhang, R.; Qian, H.; Ma, J.; Xi, N.; Cai, X.; Li, B.; Wei, D. Semantic Communication and Security over Cloud-Network-End Infrastructure: An Effective Architecture for Intelligent Mobile Systems. *IEEE Veh. Technol. Mag.* **2025**, *20*, 104–113. <https://doi.org/10.1109/MVT.2025.3568166>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.