World4V2X: A Consistency-driven World Model for Robust V2X Cooperative Perception

Rui Wang, **Shuai Wang**, **Xiangyi Qin**, **Ze Yu*** and **Xiaojun Tan***
School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University
{wangr267, qinxy36}@mail2.sysu.edu.cn, {wangsh368, yuze7, tanxj}@mail.sysu.edu.cn

Abstract

Cooperative perception allows agents to extend perceptual capabilities through inter-agent communication. However, most existing methods still adopt a frame-wise paradigm, which limits the exploitation of spatio-temporal consistency in dynamic scenes. Therefore, when observations are partially occluded or degraded, these methods are unable to leverage historical context for compensation, leading to unstable perception and reduced detection accuracy. To address these challenges, we propose **World4V2X**, the first world model framework tailored for V2X cooperative perception. The proposed method first introduces a spatial observability modeling module that defines spatial consistency boundaries to distinguish reliable regions from uncertain ones, thereby enabling spatial consistency modeling over multi-agent heterogeneous observations. Building upon this, we construct a consistency-guided world modeling module. Specifically, a temporal consistency evolution mechanism leverages features from historical and current frames to model the state evolution of the scene, capturing environmental dynamics and producing temporal consistency beliefs. Meanwhile, a consistency-guided deterministic reconstruction mechanism exploits spatial boundaries and temporal beliefs to perform diffusion-based refinement for robust cooperative perception. Extensive experiments on the OPV2V and V2XSet datasets demonstrate that World4V2X achieves state-of-the-art perception performance across diverse V2X scenarios.

1 Introduction

As autonomous driving technology continues to advance, higher levels of driving automation are imposing increasingly stringent requirements on the accuracy and robustness of perception systems [Yu *et al.*, 2025]. To meet these growing demands, cooperative perception leverages vehicle-to-everything (V2X) communication to facilitate the exchange of sensing features among multiple agents, thereby extending the effective perceptual range as well as enhancing object detection accuracy and robustness [Bai *et al.*, 2024] [Zimmer *et al.*, 2024] [Liu *et al.*, 2025]. By aggregating complementary information from diverse viewpoints, it enables vehicles to perceive objects that may be partially or even fully occluded from a single perspective [Li *et al.*, 2021] [Ren *et al.*, 2024], as shown in Figure 1-II. This capability is critical in various environments, where occlusions caused by traffic participants and adverse weather conditions can severely degrade the perception ability of individual vehicles [Wang *et al.*, 2020].

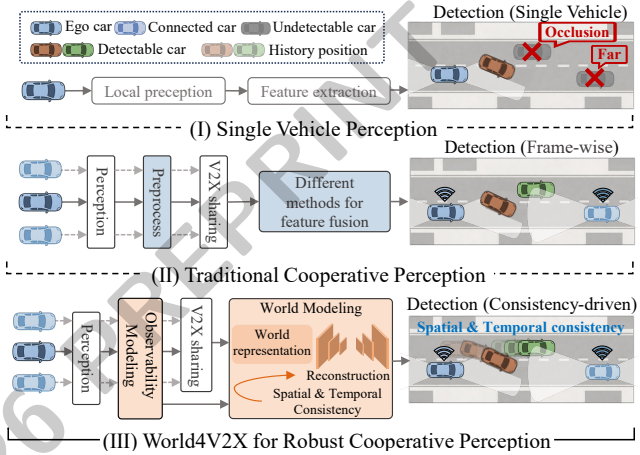


Figure 1: Comparison between World4V2X and existing pipelines. Single-vehicle perception is limited by occlusion and the sensor range. The traditional cooperative perception strategy alleviates these issues through feature exchanges, but relies solely on information from the current frame. In contrast, World4V2X adopts a consistency-driven world model paradigm. By enforcing spatial and temporal constraints, it robustly reconstructs the world state beyond instantaneous observations.

In recent years, research on V2X cooperative perception has focused primarily on improving perception accuracy [Xu *et al.*, 2022a] and reducing the incurred communication overhead [Hu *et al.*, 2022]. However, most existing approaches still adopt a frame-wise modeling paradigm, in which perception and feature fusion are performed independently at each time step. They rely totally on instantaneous sensory inputs, without modeling the spatio-temporal dependencies

*Corresponding authors.

that are inherent in dynamic traffic scenes. As a result, such frame-wise perception strategies struggle to systematically exploit the continuity of motion and the temporal evolution trend across consecutive frames. When facing challenging conditions such as severe occlusions, intermittent observation losses, or noisy sensory inputs, they often suffer from pronounced instability, leading to flickering detections and structural inconsistencies.

World models constitute a modeling paradigm with environmental state representation, temporal reasoning, and scene reconstruction capabilities, providing a promising avenue for solving the above problems. By maintaining an internal representation of the environment state, they capture temporal continuity and state transition dynamics, which in turn improve the stability and consistency of outcomes [Zuo *et al.*, 2025]. However, directly extending them to V2X cooperative perception is non-trivial. The existing world models are largely developed for single-agent settings and typically rely on centralized, ego-centric modeling assumptions [Wang *et al.*, 2024a] [Zheng *et al.*, 2025]. In contrast, V2X cooperative perception operates in a multi-agent, distributed, and heterogeneous environment. Here, observations are not only temporally sparse and intermittent but also originate from multiple agents with diverse viewpoints, sensing ranges, and sensor configurations, which substantially complicates the spatio-temporal consistency modeling and cross-agent state alignment tasks. As a result, how to construct and maintain an internal environmental representation from multi-agent observations, while leveraging spatio-temporal information to perform environment reconstruction and refinement for achieving stable and reliable cooperative perception remains a critical problem.

To address the above challenges, we integrate world modeling into V2X cooperative perception for the first time and propose a task-aligned world model, as shown in Figure 1-III. Since the behaviors of other entities are latent and uncontrollable, we focus on modeling intrinsic environmental dynamics and temporal momentum to evolve the world state, while treating state transitions as explicit changes in scene configuration. To establish the world representation, a spatial observability modeling (SOM) module is applied to characterize spatial reliability and establish a spatial consistency boundary for a subsequent consistency-guided world modeling (CWM) module. Within this module, a temporal consistency evolution (TCE) mechanism first leverages historical observations to infer temporal consistency beliefs as temporal priors. Afterward, a consistency-guided deterministic reconstruction (CDR) mechanism introduces a diffusion-based refinement process to perform targeted reconstruction by jointly exploiting the spatial boundary derived from the SOM module and temporal priors. By integrating the above proposed modules, we introduce **World4V2X**, a consistency-driven world modeling framework that treats cooperative perception as a structured state estimation task rather than a frame-wise feature fusion process. By enforcing spatial and temporal consistency in a deterministic reconstruction pipeline, World4V2X exhibits improved perception accuracy and yields a robust and interpretable alternative to conventional frame-wise cooperative perception systems.

Specifically, our contributions are summarized as follows:

1. We introduce World4V2X, to our knowledge, the first world model for V2X cooperative perception. It establishes a paradigm that prioritizes spatio-temporal consistency over frame-wise fusion, improving the adaptability of agents in diverse V2X scenarios.
2. We propose a TCE mechanism that captures environmental dynamics to generate temporal consistency beliefs, enforcing physical continuity in the evolving world representation.
3. We introduce a CDR mechanism that leverages spatio-temporal constraints for deterministic diffusion-based refinement, enabling stable and consistent world representations for cooperative perception.
4. Extensive comparative and ablation experiments on the OPV2V and V2XSet datasets demonstrate that World4V2X consistently outperforms the state-of-the-art (SOTA) methods across diverse V2X settings.

2 Related Work

2.1 Cooperative Perception

Cooperative perception has attracted widespread attention in the field of autonomous driving, with several high-quality datasets driving its development [Xu *et al.*, 2022b] [Li *et al.*, 2022] [Xu *et al.*, 2022a] [Yu *et al.*, 2022] [Xu *et al.*, 2023]. Some works have focused on improving perception accuracy through expressive fusion mechanisms, such as Transformer-based architectures like AttnFuse [Xu *et al.*, 2022b], V2X-ViT [Xu *et al.*, 2022a] and V2VFormer [Lin *et al.*, 2024]. In addition to attention-based designs, CoAlign [Lu *et al.*, 2023] introduces graph optimization to mitigate positional misalignments. TraF-Align [Song *et al.*, 2025] further addresses asynchronous perception by performing trajectory-aware feature alignment. In parallel, another line of research emphasizes the trade-off between perception performance and communication efficiency. Where2comm [Hu *et al.*, 2022] leverages confidence maps to select critical regions, whereas MRCNet [Hong *et al.*, 2024] incorporates motion-aware communication to improve robustness. CoSDH [Xu *et al.*, 2025] formulates communication as a supply–demand problem with an intermediate–late hybrid fusion strategy, and Confidence-V2X [Tan *et al.*, 2026] further employs a sparse communication strategy based on confidence-driven and information-gated approaches. However, these methods are primarily limited to frame-wise paradigms that are unable to model the spatio-temporal consistency of the environmental evolution.

2.2 World Model

A world model refers to a paradigm that possesses environmental representation, temporal awareness, and generative reconstruction capabilities [Wang *et al.*, 2025]. It constructs internal representations that are both predictive and inferable, enabling reasoning beyond immediate observations [Ding *et al.*, 2025]. Various models, including variational autoencoders (VAEs) [Abrol *et al.*, 2020], large language models (LLMs) [Vafa *et al.*, 2024], and diffusion-based generative

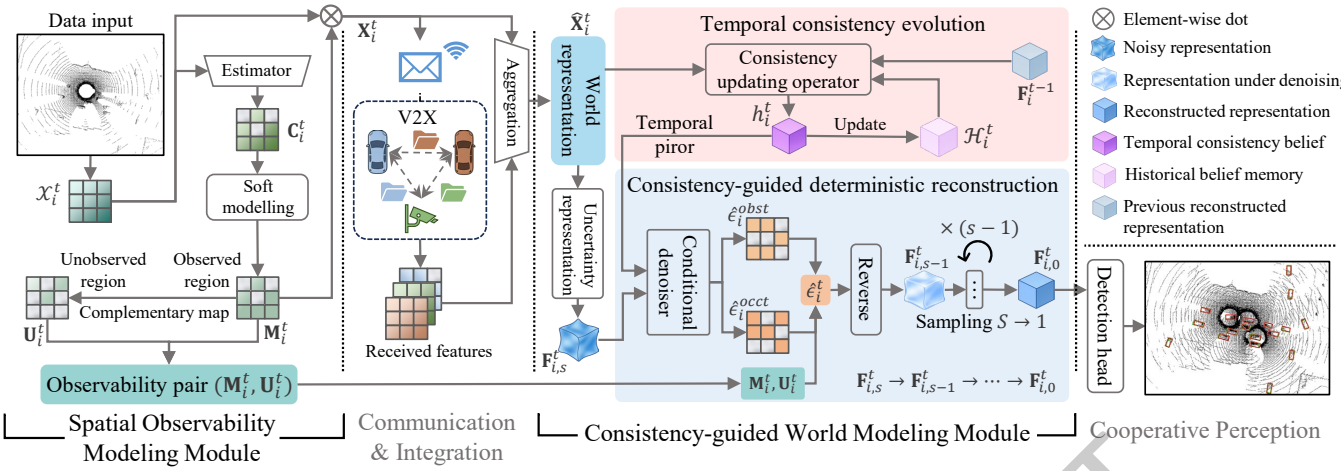


Figure 2: Overview of World4V2X. The SOM module explicitly models sensor reliability, outputting a spatial observability pair $(\mathbf{M}_i^t, \mathbf{U}_i^t)$ that serves as the spatial consistency boundary. The TCE mechanism captures temporally latent dynamics to generate a temporal consistency belief $\mathbf{h}_i^t$ by a consistency updating operator Φ. The CDR mechanism integrates the spatial boundary and temporal belief as consistency constraints to guide the deterministic refinement of the world representation for cooperative perception.

models [Wang *et al.*, 2024a], have been widely used to construct world models across different application domains, including autonomous driving [Guan *et al.*, 2024]. For example, DriveDreamer-2 [Zhao *et al.*, 2025] explores controllable future scene generation tasks conditioned on textual guidance by applying an LLM. World4Drive [Zheng *et al.*, 2025] incorporates semantic priors derived from foundation vision models to support multimodal trajectory reasoning work. DriveWM [Wang *et al.*, 2024b] focuses on spatio-temporal latent modeling for multi-view future prediction tasks, and LAW [Li *et al.*, 2025] employs self-supervised latent dynamics to enhance end-to-end driving decisions with reduced reliance on costly perception annotations. In contrast with the existing world models that were primarily designed for single-agent settings, the proposed World4V2X is a task-aligned world model that targets cooperative perception.

3 Methodology

In this section, we provide a detailed introduction to the algorithmic workflow of World4V2X.

3.1 Overview

The overall framework of World4V2X is illustrated in Figure 2. At each time step, each agent first applies the SOM module to explicitly identify reliable and uncertain regions, establishing a spatial consistency boundary. After V2X communication is performed, the shared features are aggregated into a unified world representation, which is then processed by the CWM module. Within this module, a TCE mechanism infers a temporal consistency belief from the historical observations, providing a temporal prior over the current scene. Subsequently, a CDR mechanism refines the noisy world representation through a deterministic diffusion-based process under the joint guidance of spatial and temporal consistency. The final reconstructed world representation is fed into the detection head for performing 3D object detection.

3.2 Spatial Observability Modeling Module

Given a time t, agent i captures its raw perceptual data and extracts the result as a BEV feature map $\mathcal{X}_i^t \in \mathbb{R}^{H\times W\times D}$, where D denotes the number of feature channels, and H and W represent the spatial dimensions.

Modeling spatial observability is critical for establishing spatial consistency, as it determines whether a region is reliable or uncertain for subsequent world modeling. To achieve this, we predict a dense confidence map:

$$\mathbf{C}_i^t = \sigma\left(\varepsilon\left(\mathcal{X}_i^t\right)\right) \in [0,1]^{H\times W}, \tag{1}$$

where $\varepsilon(\cdot)$ is a lightweight estimator achieved by a 2D convolution and $\sigma(\cdot)$ denotes the sigmoid function. This map quantifies the local reliability of the spatial structure.

Rather than applying a hard thresholding scheme that completely discards uncertain regions and destroys structural continuity [Hu *et al.*, 2022], we use a soft modelling strategy. The raw confidence predictions are transformed into an observability map $\mathbf{M}_i^t \in [0,1]^{H\times W}$, which is applied to the BEV feature map to obtain reliable observations:

$$\mathbf{X}_i^t = \mathcal{X}_i^t \odot \mathbf{M}_i^t = \mathcal{X}_i^t \odot \mathcal{S}\left(\mathbf{C}_i^t\right), \tag{2}$$

where $\mathcal{S}(\cdot)$ denotes the operator for performing normalization and scaling, which can enhance the relative confidence difference while preventing mask degradation. $\odot$ denotes the element-wise multiplication operation.

A spatial observability pair $(\mathbf{M}_i^t, \mathbf{U}_i^t)$ is generated as a spatial consistency boundary, where $\mathbf{U}_i^t = \mathbf{1} - \mathbf{M}_i^t$ highlights unobserved regions. This pair provides spatial consistency for the subsequent world modeling, as $\mathbf{M}_i^t$ enforces alignment with the sensing results obtained in the observed areas, and $\mathbf{U}_i^t$ indicates the need for the temporal consistency beliefs generated by the TCE mechanism to be reconstructed. By explicitly modeling this boundary, the reconstructed world representation is still based on reliable observation, while maintaining spatial consistency with the generated reconstruction.

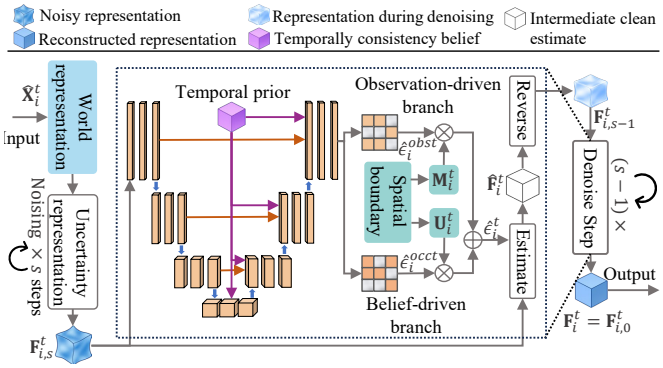


Figure 3: Overview of the CDR mechanism. It uses temporal consistency beliefs as temporal priors to guide the diffusion-based refinement process.

Finally, the resulting preprocessed feature $\mathbf{X}_i^t$ is treated as a reliable observation, which is sent to other agents during V2X communication.

3.3 World Modeling with Consistency Guidance

After performing V2X communication, agent i aggregates the features received from its neighbors on its own features to form a collaborative observation $\hat{\mathbf{X}}_i^t$. We treat it as the world representation at time t, which preserves the spatial structure required for cooperative perception while serving as the inference target for the subsequent reconstruction process.

In practical V2X systems, the collaborative world representation inevitably contains uncertainty caused by the residual localization errors and communication delays induced across different agents. To consider the impact of these uncertainties, we employ an uncertainty representation module to obtain the noisy world state $\mathbf{F}_{i,s}^t$ by sampling a noise timestep $s \sim \{1, 2, \ldots, S\}$:

$$\mathbf{F}_{i,s}^t = \sqrt{\overline{\alpha}_s} \cdot \hat{\mathbf{X}}_i^t + \sqrt{1 - \overline{\alpha}_s} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

where $\overline{\alpha}_s$ defines the cumulative noise schedule at step s. $\epsilon \sim \mathcal{N}(0, I)$ denotes standard Gaussian noise, which facilitates learning under uncertainty instead of contributing to the semantics of the inferred world state.

Here, $\mathbf{F}_{i,s}^t$ serves as a common inference scaffold for the subsequent stages, where it is deterministically reconstructed under the joint guidance of spatial consistency boundaries and temporal consistency beliefs.

Temporal Consistency Evolution

Instead of performing a frame-wise process independently at each time step, we exploit temporal consistency to capture the evolution of the environment state, thereby supporting inferences beyond instantaneous observations.

To achieve this, agent i is required to update a temporal belief memory $\mathcal{H}_i^t$ at each frame, which encapsulates the historical evolution of the scene. This memory serves as a compact internal state that accumulates long-term temporal context and provides a stable reference for evaluating newly observed features.

Building upon it, we integrate information from multiple sources to infer a temporal consistency belief $\mathbf{h}_i^t$. Specifically, at time t, three complementary inputs are taken: the previous temporal belief memory $\mathcal{H}_i^{t-1}$, the current collaborative observation $\hat{\mathbf{X}}_i^t$, which provides instantaneous multi-agent sensing information, and the previously reconstructed world representation $\mathbf{F}_i^{t-1}$, which represents the world state derived from the previous frame.

On the basis of these inputs, a temporal consistency belief $\mathbf{h}_i^t$ is inferred through a learnable consistency updating operator:

$$\mathbf{h}_i^t = \Phi\left(\mathcal{H}_i^{t-1}, \hat{\mathbf{X}}_i^t, \mathbf{F}_i^{t-1}\right), \tag{4}$$

where $\Phi(\cdot)$ is implemented as a lightweight fusion network that is composed of stacked convolutional layers with a gating mechanism. It aligns the historical beliefs with the current observation and the previously reconstructed state, which performs two critical functions. First, by jointly analyzing $\mathcal{H}_i^{t-1}$ and $\mathbf{F}_i^{t-1}$, it aggregates historical information to form a temporal consistency belief regarding the current scene. Second, by aligning the temporal belief with the incoming observation $\hat{\mathbf{X}}_i^t$, it serves as a form of state transition that corrects the accumulated drift and suppresses inconsistent short-term fluctuations.

After the belief $\mathbf{h}_i^t$ is obtained, the temporal belief memory is updated through a recurrent mechanism to prepare for the next time step:

$$\mathcal{H}_i^t = \text{Update}(\mathcal{H}_i^{t-1}, \mathbf{h}_i^t), \tag{5}$$

where $\text{Update}(\cdot)$ is implemented by a multi-layer spatial GRU to selectively retain long-term historical context while incorporating new information.

The resulting belief $\mathbf{h}_i^t$ represents a temporal consistency-aware hypothesis concerning the current world state. Unlike raw or instantaneous observations, it encodes temporally coherent scene semantics and serves as a temporal prior for the subsequent reconstruction module, guiding stable and consistent world refinement.

Consistency-Guided Deterministic Reconstruction

The workflow of this module is illustrated in Figure 3, which deterministically recovers a coherent world state under spatio-temporal consistency constraints, rather than performing generative synthesis or stochastic imagination. Instead of synthesizing representations from scratch, we approach this process as a refinement task, aiming to preserve observation-supported structures while correcting inconsistencies.

Unlike the stochastic sampling procedures of generative models that introduce diversity, we formulate reconstruction as a deterministic diffusion-based refinement process, inspired by DDIM [Song *et al.*, 2020], where the reverse trajectory is fully constrained by spatial and temporal priors, ensuring stable and physically consistent reconstruction for V2X systems.

Given the noisy representation $\mathbf{F}_{i,s}^t$, during inference, the conditional denoiser predicts two noise components using a shared network:

$$\left(\hat{\epsilon}_i^{obst}, \hat{\epsilon}_i^{occt}\right) = \epsilon_\theta\left(\mathbf{F}_{i,s}^t, \mathbf{h}_i^t, s\right), \tag{6}$$

where $\hat{\epsilon}_i^{obst}$ estimates the noise in reliable regions, and $\hat{\epsilon}_i^{occt}$ targets unobserved areas that require inference guided by temporal consistency. These noise components are spatially fused to enforce the dual consistency constraints:

$$\hat{\epsilon}_i^t = \mathbf{M}_i^t \odot \hat{\epsilon}_i^{obst} + \mathbf{U}_i^t \odot \hat{\epsilon}_i^{occt}, \tag{7}$$

where $\odot$ denotes the element-wise multiplication operation. This asymmetric conditional inference scheme ensures that reconstruction is observation-driven in well-observed regions and belief-driven in occluded areas, with both processes unified within a single deterministic refinement step.

Following the diffusion standard formulation, we first estimate the noise-free world representation $\hat{\mathbf{F}}_i^t$:

$$\hat{\mathbf{F}}_i^t = \frac{\mathbf{F}_{i,s}^t - \sqrt{1-\overline{\alpha}_s}\,\hat{\epsilon}_i^t}{\sqrt{\overline{\alpha}_s}}, \tag{8}$$

which serves as the intermediate clean estimate. Then, $\mathbf{F}_{i,s-1}^t$ is computed through the reverse schedule:

$$\mathbf{F}_{i,s-1}^t = \text{ReverseStep}(\mathbf{F}_{i,s}^t, s, \hat{\epsilon}_i^t), \quad s = S \to 1. \tag{9}$$

As the denoising process progresses, the final reconstructed world $\mathbf{F}_i^t = \mathbf{F}_{i,0}^t$ achieves a unified consistency, which is spatially anchored to the available sensor evidence and temporally coherent with the historical evolution of the scene. This representation is fed into the detection head for 3D object detection.

3.4 Training Loss

World4V2X is entirely driven by the perception objective, without requiring any ground-truth supervision. Thus, it learns to reconstruct the world state only when it really benefits the downstream detection task. This design ensures that the reconstructed world state is temporally consistent and task-oriented, rather than behaving as an unconstrained generative prior.

Given the reconstructed BEV feature $\hat{\mathbf{F}}_i^t$, our detection head predicts object heatmaps and 3D box parameters. We directly train World4V2X in an end-to-end manner by adapting the standard CenterPoint loss formulation [Yin *et al.*, 2021], where a Gaussian focal loss supervises the heatmap classification process, and a Smooth-L1 regression loss is applied to the offsets, height, dimensions and orientation.

4 Experiments

We evaluate the proposed algorithm using the widely adopted object detection task, with comprehensive experiments conducted on two public datasets, i.e., OPV2V [Xu *et al.*, 2022b] and V2XSet [Xu *et al.*, 2022a], under both synchronous (sync) and asynchronous (async) settings.

4.1 Experimental Setup

Datasets

OPV2V is the first large-scale simulated V2V dataset generated by OpenCDA [Xu *et al.*, 2021] and CARLA [Dosovitskiy *et al.*, 2017]. It consists of 11,464 LiDAR-based multi-agent frames covering 73 scenarios in 9 cities with diverse road structures. The official split includes 6764, 1981, 2719 frames for training, validation, and testing.

V2XSet is a CARLA-based dataset for V2X perception tasks under realistic communication constraints, including localization noise and transmission delays. It contains 11,447 multi-agent LiDAR frames, with 6694, 1920, and 2833 frames for training, validation, and testing, respectively.

Evaluation Metrics

To evaluate the performance of the tested algorithms, we compare the object detection accuracies achieved for vehicles under cooperative perception, which are measured with the average precision (AP) values produced at intersection-over-union (IoU) thresholds of 0.5 (AP@50) and 0.7 (AP@70).

Implementation details

To comprehensively validate the performance of the algorithms, the environment is divided into sync and async modes. In the sync mode, there is no communication delay or pose misalignment, and each algorithm is evaluated under ideal conditions. In the async mode, we set the transmission delay to 200 ms, and the latency of the backbone calculation to 10 ms. Gaussian noise is added to simulate the pose errors of non-ego vehicles: $N(0, \sigma_t^2)$ on the 2D centers of the true global pose (x, y), and $N(0, \sigma_r^2)$ on its yaw angle θ_p. The parameter for both types is set to 0.2.

During the training phase, an agent is randomly selected as the ego car. When testing, a fixed ego car is used to evaluate all models. The communication range is set to 70 m, and the number of connected vehicles is limited to 5. PointPillars [Lang *et al.*, 2019] is used as the feature encoder, and the voxel size is set to $(0.4m, 0.4m, 4m)$, capping the maximum number of points per voxel at 32.

Given that computational efficiency is critical for real-time V2X systems, we set the number of denoising steps $S = 2$ in the CDR mechanism by considering the trade-off between algorithmic performance and computational cost. We use the Adam optimizer with a learning rate of 0.001 and a gradual weight decay factor of 0.0001. The batch size is set to 4. The proposed model is trained using an A100 GPU.

Comparison Methods

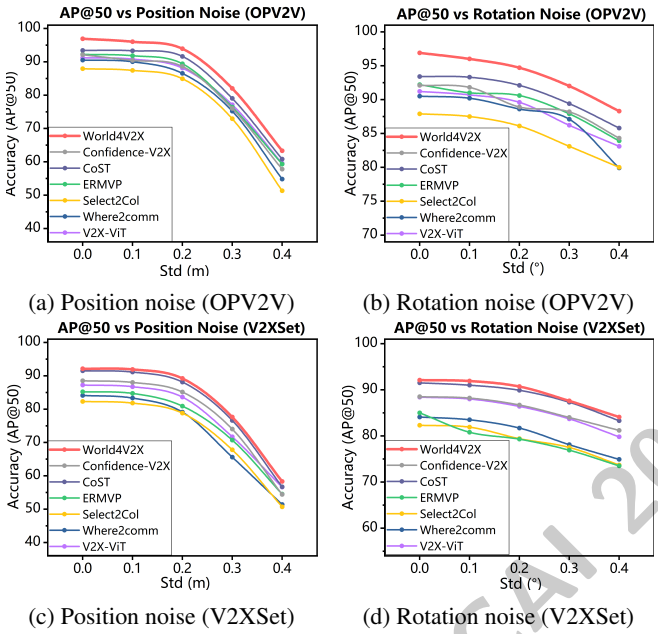
We compare World4V2X with single car perception and 6 SOTA methods: V2X-ViT [Xu *et al.*, 2022a], Where2comm [Hu *et al.*, 2022], Select2Col [Liu *et al.*, 2024], ERMVP [Zhang *et al.*, 2024], CoST [Tang *et al.*, 2025], and Confidence-V2X [Tan *et al.*, 2026].

4.2 Quantitative results

Object detection performance. Table 1 shows the 3D detection results produced on 2 datasets under different environments. On the OPV2V dataset, in the sync setting, World4V2X achieves an AP@50 score of 96.9, surpassing the second-best model, CoST, by 3.5% (93.4). At the same time, its AP@70 score of 92.8 exceeds the second-best value of 88.9 by 3.9%. In the async setting, World4V2X continues to demonstrate its dominance by achieving AP@50 and AP@70 scores of 95.6 and 85.3, respectively, which are 3.3% and 3.2% higher than the corresponding second-best scores.

Method	Publications	OPV2V				V2XSet			
		Sync		Async		Sync		Async	
		AP@50	AP@70	AP@50	AP@70	AP@50	AP@70	AP@50	AP@70
Single car	–	67.9	60.2	67.9	60.2	61.3	44.7	61.3	44.7
V2X-ViT [Xu *et al.*, 2022a]	ECCV 2022	91.2	83.3	90.0	75.2	87.2	75.6	85.2	60.6
Where2comm [Hu *et al.*, 2022]	NIPS 2022	90.5	84.2	88.7	74.1	84.1	70.7	82.9	61.1
Select2Col [Liu *et al.*, 2024]	TVT 2024	87.9	79.7	85.8	72.3	82.3	72.5	79.5	62.4
ERMVP [Zhang *et al.*, 2024]	CVPR 2024	92.2	85.6	89.8	82.1	85.2	74.7	81.1	65.9
CoST [Tang *et al.*, 2025]	ICCV 2025	93.4	88.9	92.3	80.8	91.5	81.4	88.2	72.3
Confidence-V2X [Tan *et al.*, 2026]	AEI 2026	92.1	85.2	89.7	79.5	88.5	80.0	85.5	70.3
World4V2X (Ours)	–	**96.9**	**92.8**	**95.6**	**85.3**	**92.1**	**84.7**	**91.6**	**78.7**

Table 1: Performance achieved on the OPV2V and V2XSet datasets under sync and async settings.



(a) Position noise (OPV2V) (b) Rotation noise (OPV2V)

(c) Position noise (V2XSet) (d) Rotation noise (V2XSet)

Figure 4: Performance achieved on the OPV2V and V2XSet datasets under various noise settings.

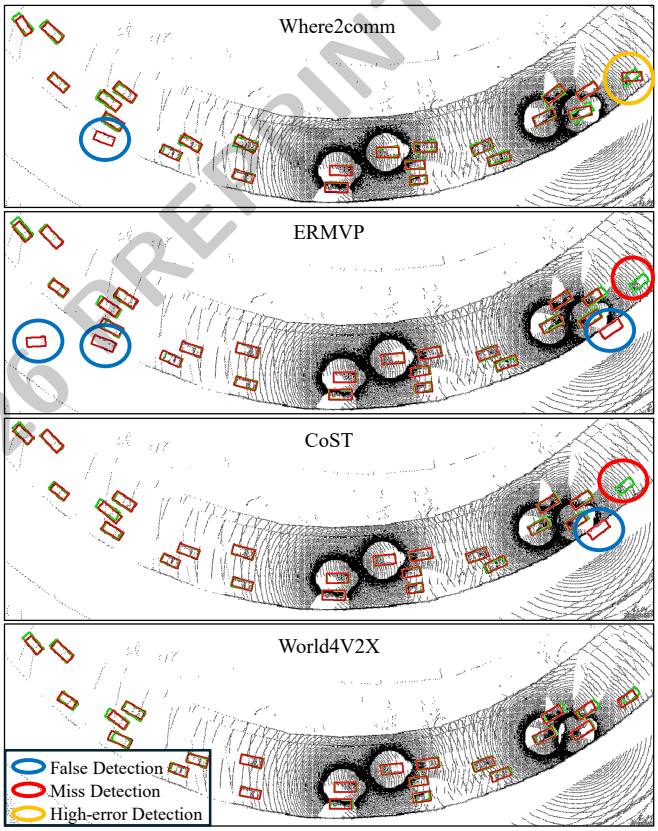


Figure 5: Visualization of the detection results produced for a bend on the OPV2V dataset. Owing to the spatio-temporal consistency constraints within World4V2X, its detection results are more stable than those of other methods, avoiding abrupt variations in the detection outputs produced across consecutive frames.

Similar results can be drawn from the V2XSet dataset, which shows that World4V2X achieves superior performance.

Robustness to position and rotation errors. To evaluate the consistency of the proposed approach, in the async setting, we vary the noise levels to assess the sensitivity of the tested methods to position and rotation errors. As shown in Figure 4, when the noise level increases, the performance of all the models gradually decreases. However, World4V2X consistently maintains its lead, demonstrating strong robustness in complex environments. This robustness can be attributed to two factors. First, the TCE mechanism captures environmental evolution dynamics, which helps correct local misalignments by enforcing physical continuity. Second, the CDR mechanism conditions the refinement process on temporal consistency beliefs, enabling the model to filter out stochastic inconsistencies caused by localization and rotation errors induced during the inference process.

4.3 Qualitative Results

Visualization of the object detection results. We randomly select a scene to demonstrate the detection results produced by Where2comm, ERMVP, CoST, and World4V2X. As shown in Figure 5, the other methods exhibit significant

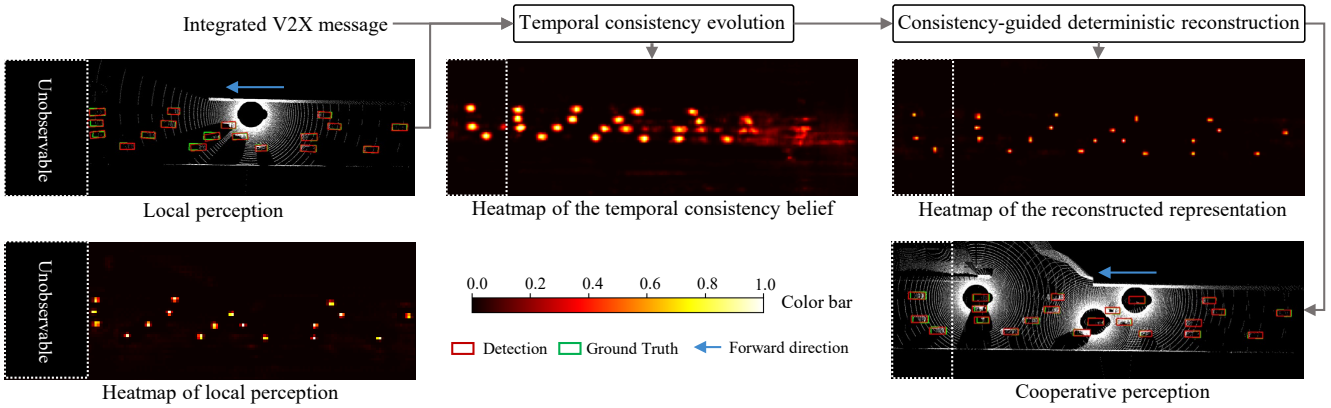


Figure 6: Heatmaps generated by World4V2X. The central heatmap displays the temporal consistency belief distribution generated by the TCE mechanism. A dynamic attention shift is observable: as the car moves left, the belief intensity (bright yellow) actively shifts toward the forward direction to anticipate potential risks, even in regions that can not be unobserved by local sensors. Conversely, the attention given to the traversed rear region naturally decreases (darker red).

#	Component			Sync		Async	
	SOM	TCM	CDR	AP@50	AP@70	AP@50	AP@70
0	✓	✓	✓	**92.1**	**84.7**	**91.6**	**78.7**
1		✓	✓	89.8	80.5	89.0	71.5
2	✓		✓	90.7	79.8	90.0	70.1
3	✓			87.8	67.2	86.7	63.3

Table 2: Ablation studies for assessing the contribution of the main model components on the V2XSet dataset.

biases, including false detections (detecting non-existent objects), missed detections (failing to detect actual objects), and high-error detections (low degrees of overlap between the predicted and actual bounding boxes). Owing to its ability to maintain spatial and temporal consistency, World4V2X provides more precise and comprehensive perception capabilities. This is evidenced by the fact that its predicted bounding boxes closely match the actual targets.

Visualization of heatmaps. We visualize the attention distribution produced by World4V2X in a challenging occlusion scenario. As shown in Figure 6, while the local perception ability of the ego car is limited by its range, the TCE mechanism successfully integrates historical dynamics with V2X messages to infer a global temporal consistency belief. Since the vehicle is moving to the left, the algorithm generates a belief that intelligently concentrates high confidence (indicated by the bright yellow clusters) in the forward direction. This highlights the critical objects located in the future path of the vehicle, even those that are outside of its sensing range. Simultaneously, the attention given to the regions that have already traversed behind the ego vehicle gradually decreases (shown in darker shades), as these areas pose less risk to the current process. This temporal belief acts as a semantic spotlight, guiding the subsequent reconstruction module to efficiently allocate its generative capacity.

4.4 Ablation Studies

We investigate the contribution of each core component contained in World4V2X by comparing the full framework (Model 0) with three architectural variants. Model 1 removes the SOM module. This system lacks a spatial consistency boundary for distinguishing between reliable and uncertain regions. The noise estimation task is instead treated uniformly during the reconstruction process. Model 2 excludes the TCE mechanism. Its reconstruction process lacks the guidance of temporal priors and relies solely on the current spatial context without leveraging historical dynamics for stabilizing the refinement. Model 3 removes the TCE and CDR mechanisms, serving as the baseline for frame-wise fusion. It directly feeds the max-pooling aggregated features into the detection head without the reconstruction process.

As shown in Table 2, Model 0 achieves the best performance, validating that robust cooperative perception arises from the interaction among all three components. Removing any module leads to a notable performance degradation, indicating that each component plays a vital role.

5 Conclusion

This paper introduces World4V2X, the first world modeling framework for achieving robust V2X cooperative perception. Unlike the traditional approaches, World4V2X advances the paradigm from frame-wise fusion to a task-aligned world model that is driven by spatio-temporal consistency. By integrating spatial observability modeling with consistency-guided world modeling, by means of its TCE and CDR mechanisms, it enables robust and stable cooperative perception beyond instantaneous observations. Extensive experiments conducted on two datasets demonstrate that World4V2X consistently outperforms the SOTA methods across diverse and challenging environments. In the future, we plan to explore the potential of this world model framework for use in downstream planning tasks, as well as real-world deployments implemented under varying communication bandwidths.

Acknowledgments

The research is supported by Shenzhen Science and Technology Program (Grant No.ZDCY20250901112210012 & KJZD20231023100204008).

References

[Abrol *et al.*, 2020] Vinayak Abrol, Pulkit Sharma, and Arijit Patra. Improving generative modelling in vaes using multimodal prior. *IEEE Transactions on Multimedia*, 23:2153–2161, 2020.

[Bai *et al.*, 2024] Zhengwei Bai, Guoyuan Wu, Matthew J Barth, Yongkang Liu, Emrah Akin Sisbot, Kentaro Oguchi, and Zhitong Huang. A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Ding *et al.*, 2025] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

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

[Guan *et al.*, 2024] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

[Hong *et al.*, 2024] Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. Multi-agent collaborative perception via motion-aware robust communication network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15301–15310, 2024.

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

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

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

[Li *et al.*, 2022] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022.

[Li *et al.*, 2025] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Lin *et al.*, 2024] Chunmian Lin, Daxin Tian, Xuting Duan, Jianshan Zhou, Dezong Zhao, and Dongpu Cao. V2vformer: Vehicle-to-vehicle cooperative perception with spatial-channel transformer. *IEEE Transactions on Intelligent Vehicles*, 9(2):3384–3395, 2024.

[Liu *et al.*, 2024] Yuntao Liu, Qian Huang, Rongpeng Li, Xianfu Chen, Zhifeng Zhao, Shuyuan Zhao, Yongdong Zhu, and Honggang Zhang. Select2col: Leveraging spatial-temporal importance of semantic information for efficient collaborative perception. *IEEE Transactions on Vehicular Technology*, 2024.

[Liu *et al.*, 2025] Haizhuang Liu, Huazhen Chu, Junbao Zhuo, Bochao Zou, Jiansheng Chen, and Huimin Ma. Sparsecomm: An efficient sparse communication framework for vehicle-infrastructure cooperative 3d detection. *Pattern Recognition*, 158:110961, 2025.

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

[Ren *et al.*, 2024] Shunli Ren, Zixing Lei, Zi Wang, Mehrdad Dianati, Yafei Wang, Siheng Chen, and Wenjun Zhang. Interruption-aware cooperative perception for v2x communication-aided autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020.

[Song *et al.*, 2025] Zhiying Song, Lei Yang, Fuxi Wen, and Jun Li. Traf-align: Trajectory-aware feature alignment for asynchronous multi-agent perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12048–12057, 2025.

[Tan *et al.*, 2026] Xiaojun Tan, Rui Wang, Jinping Wang, Shuai Wang, Xu Wang, and Dongsheng Wu. Confidence-v2x: Confidence-driven sparse communication for efficient v2x cooperative perception. *Advanced Engineering Informatics*, 69:103914, 2026.

[Tang *et al.*, 2025] Zongheng Tang, Yi Liu, Yifan Sun, Yulu Gao, Jinyu Chen, Runsheng Xu, and Si Liu. Cost: Efficient collaborative perception from unified spatiotemporal perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1120–1129, October 2025.

[Vafa *et al.*, 2024] Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In

A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 26941–26975. Curran Associates, Inc., 2024.

[Wang *et al.*, 2020] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European conference on computer vision*, pages 605–621. Springer, 2020.

[Wang *et al.*, 2024a] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024.

[Wang *et al.*, 2024b] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[Wang *et al.*, 2025] Jinping Wang, Weiwei Song, Hao Chen, Jinchang Ren, and Huimin Zhao. Fusdreamer: Label-efficient remote sensing world model for multimodal data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63(3), 2025.

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

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

[Xu *et al.*, 2022b] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022.

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

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

[Yin *et al.*, 2021] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

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

[Yu *et al.*, 2025] Ze Yu, Jun Li, Zesong Chen, Yuzhen Wei, Xiaofei Zhang, and Xiaojun Tan. Multimodal end-to-end autonomous driving via bilateral modality interaction. *Expert Systems with Applications*, page 128458, 2025.

[Zhang *et al.*, 2024] Jingyu Zhang, Kun Yang, Yilei Wang, Hanqi Wang, Peng Sun, and Liang Song. Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12575–12584, June 2024.

[Zhao *et al.*, 2025] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10412–10420, 2025.

[Zheng *et al.*, 2025] Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28632–28642, 2025.

[Zimmer *et al.*, 2024] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. Tumtraf v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22668–22677, 2024.

[Zuo *et al.*, 2025] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6772–6781, 2025.