# V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer

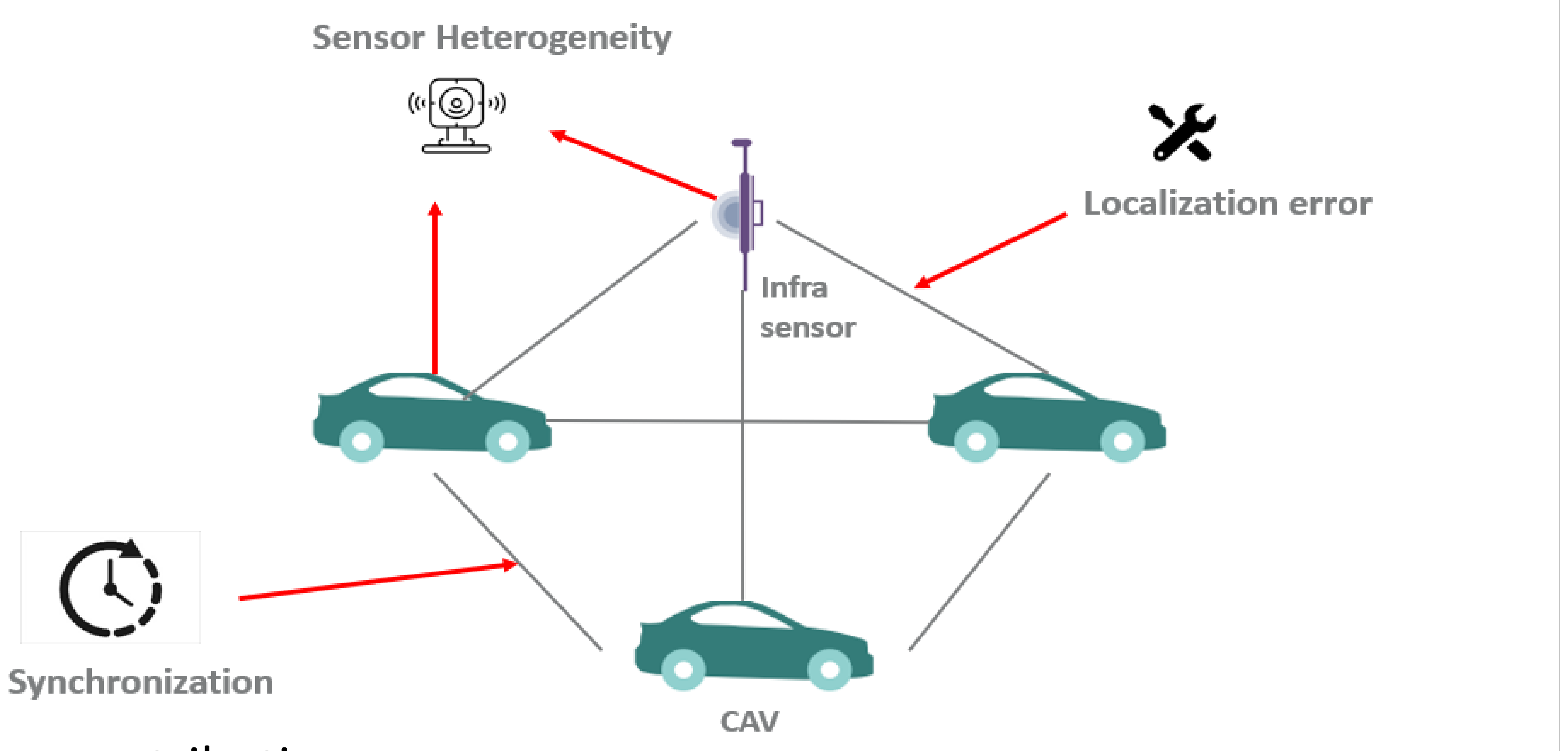
Runsheng Xu[1*], Hao Xiang[1*], Zhengzhong Tu[2*], Xin Xia[1], Ming-Hsuan Yang[3,4] , Jiaqi Ma[1]

[1]UCLA & [2]UT-Austin& [3]Google Research & [4]UC Merced

* Indicates equal contribution

**Paper, code, and data are available:**
https://github.com/DerrickXuNu/v2x-vit

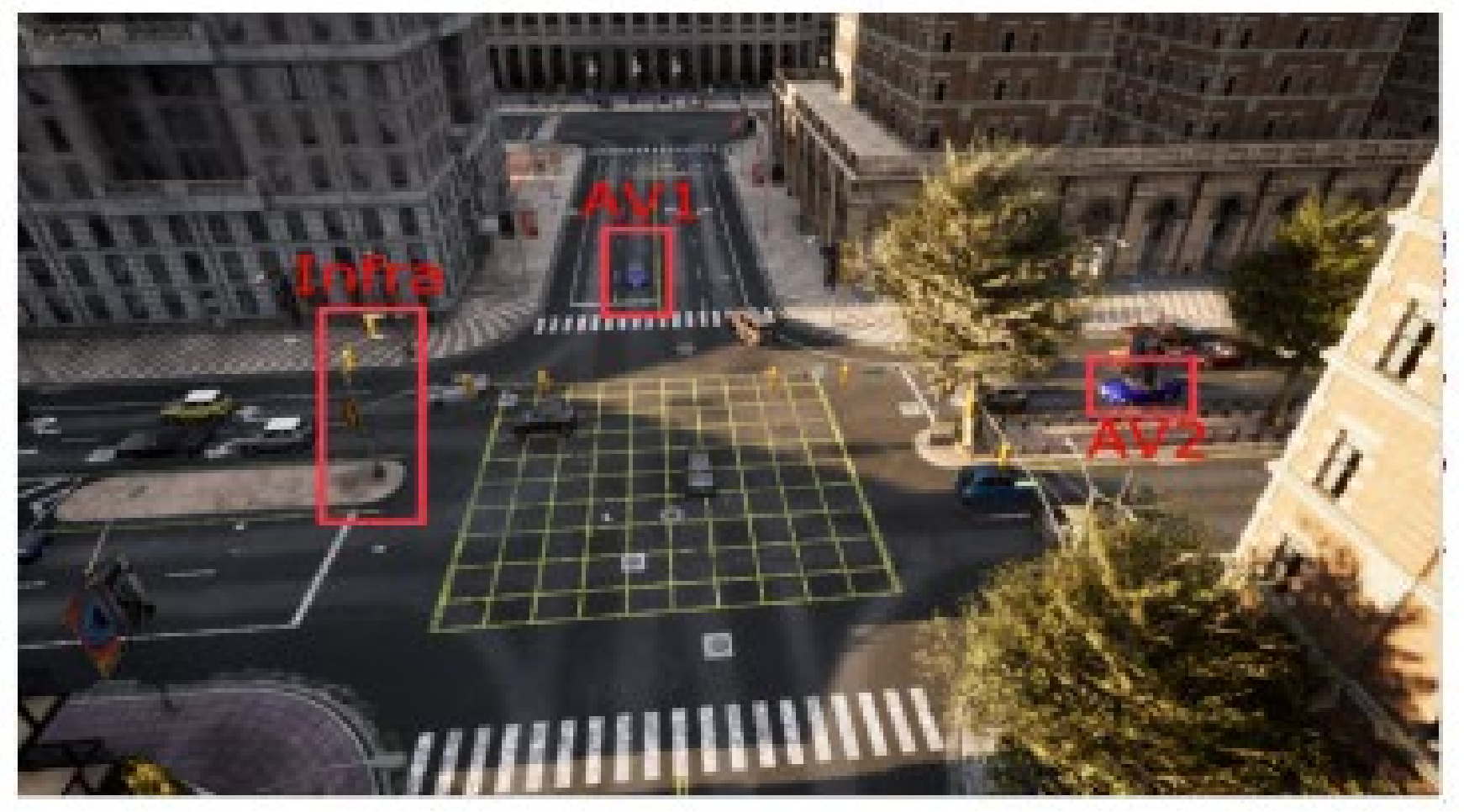
## V2X Cooperative Challenge



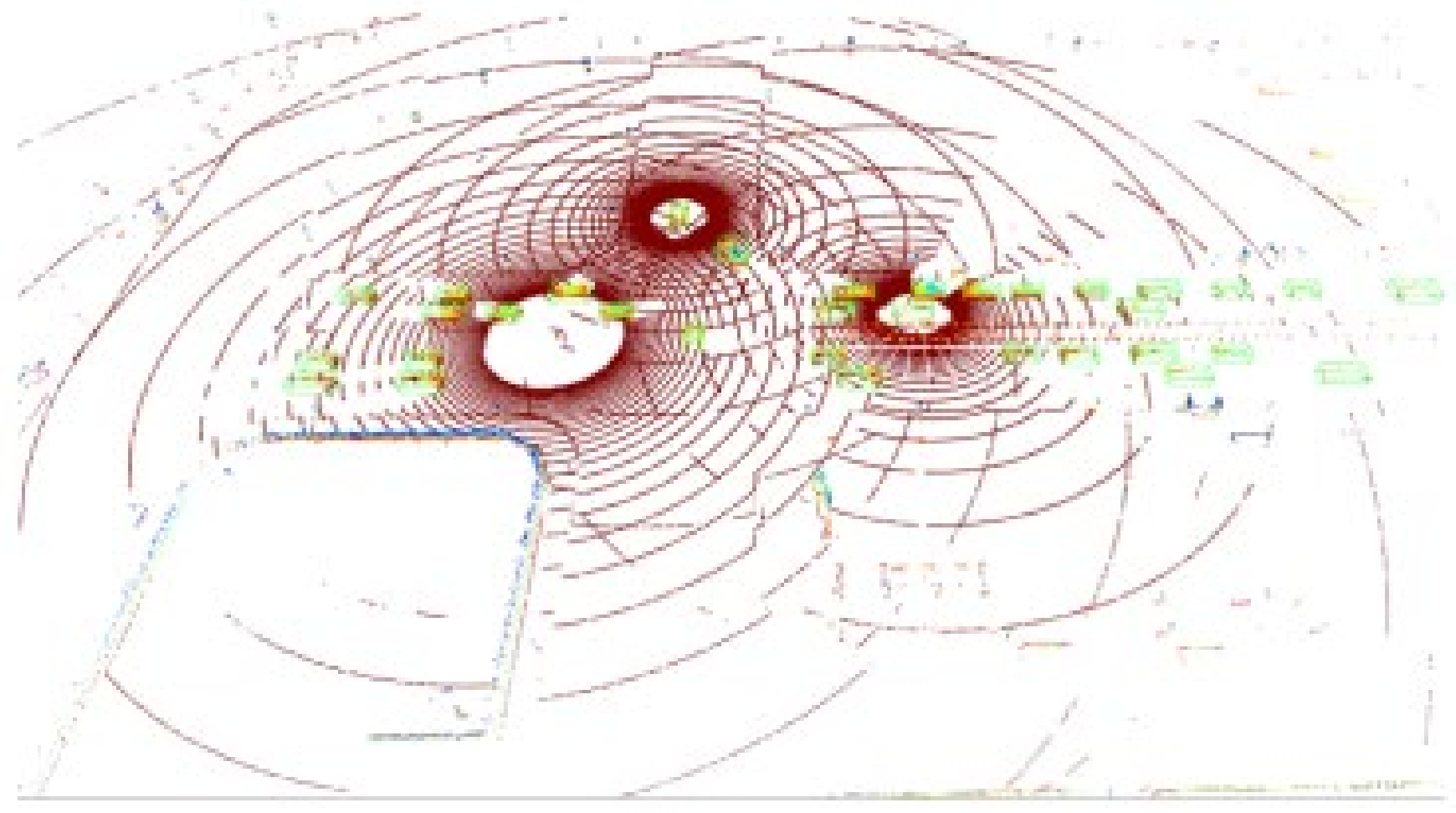
Our contributions:

- We present the first unified transformer architecture (V2X-ViT) for V2X perception, which can capture the heterogeneity nature of V2X systems with strong robustness against various noises.
- We propose a novel heterogeneous multi-agent attention module (HMSA) tailored for adaptive information fusion between heterogeneous agents.
- We present a new multi-scale window attention module (MSwin) that simultaneously captures local and global spatial feature interactions in parallel.
- We construct V2XSet, a new large-scale open simulation dataset for V2X perception, which explicitly accounts for imperfect real-world conditions.

## V2XSet: A new V2X Perception dataset

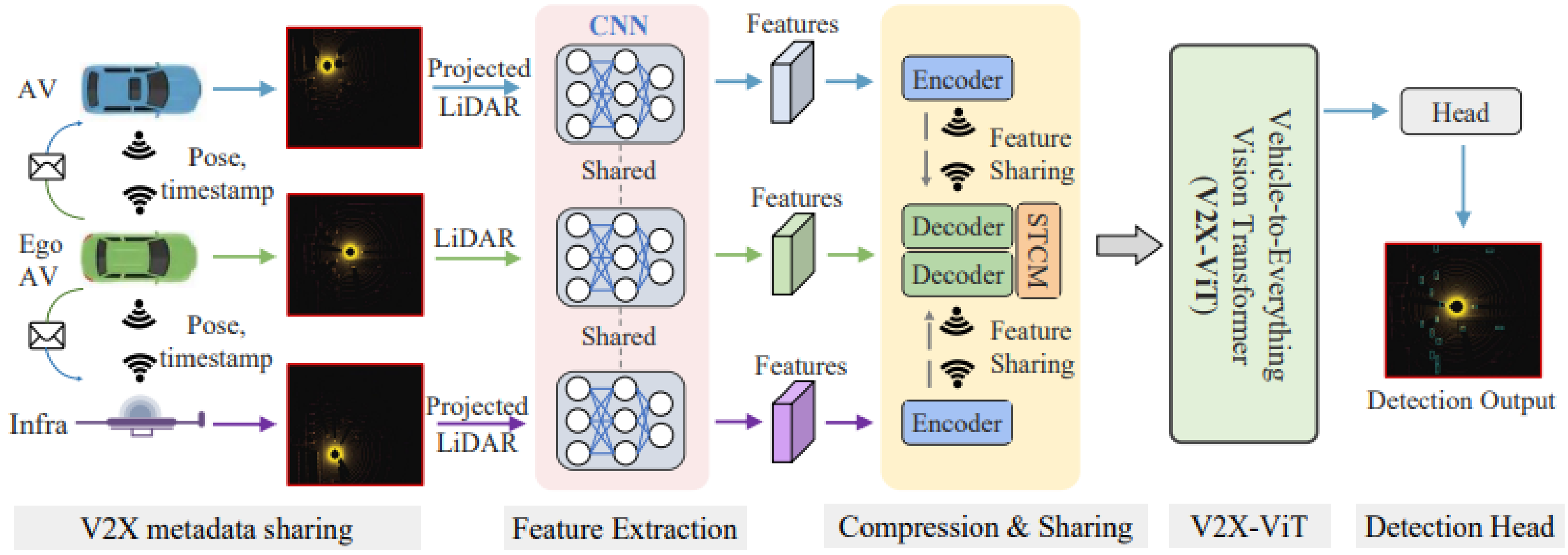


(a) Snapshot of Simulation



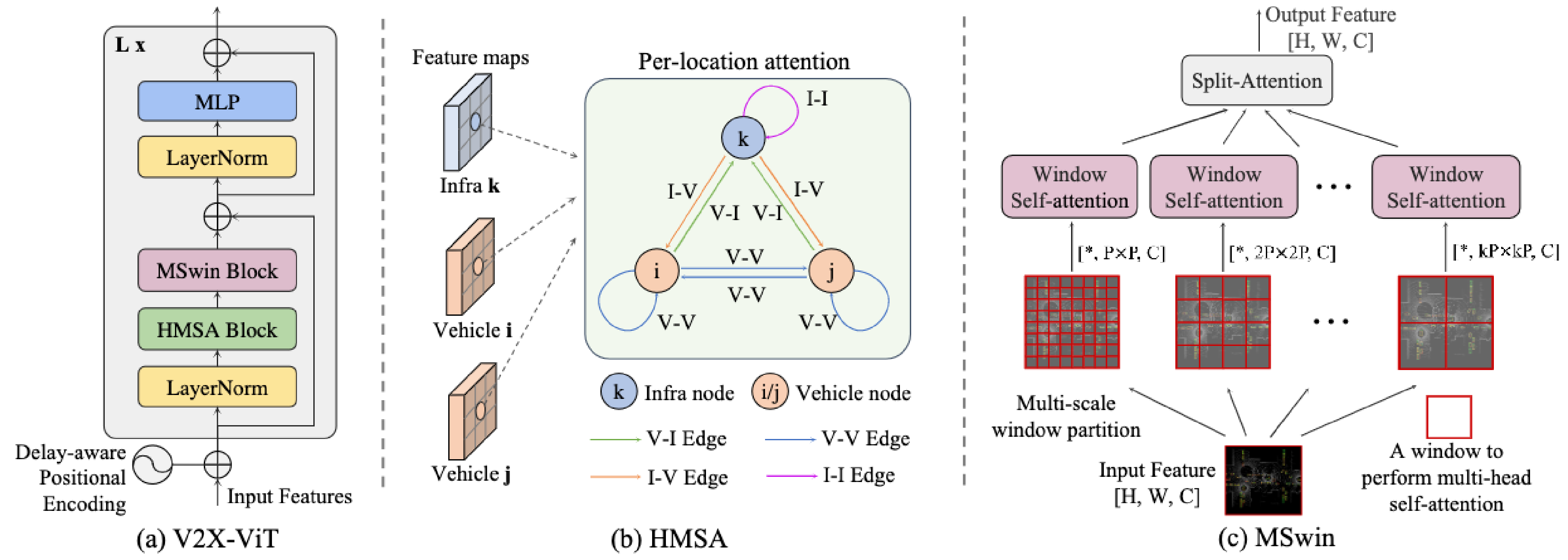
(b) Aggregated LiDAR point cloud

## V2X-ViT Overall Framework



## V2X-ViT Architecture

- Learn inter-agent interaction and per-agent spatial attention.
- HMSA captures heterogeneity between infra and vehicle.
- Mswin improves the robustness against localization error
- DPE encodes the temporal information

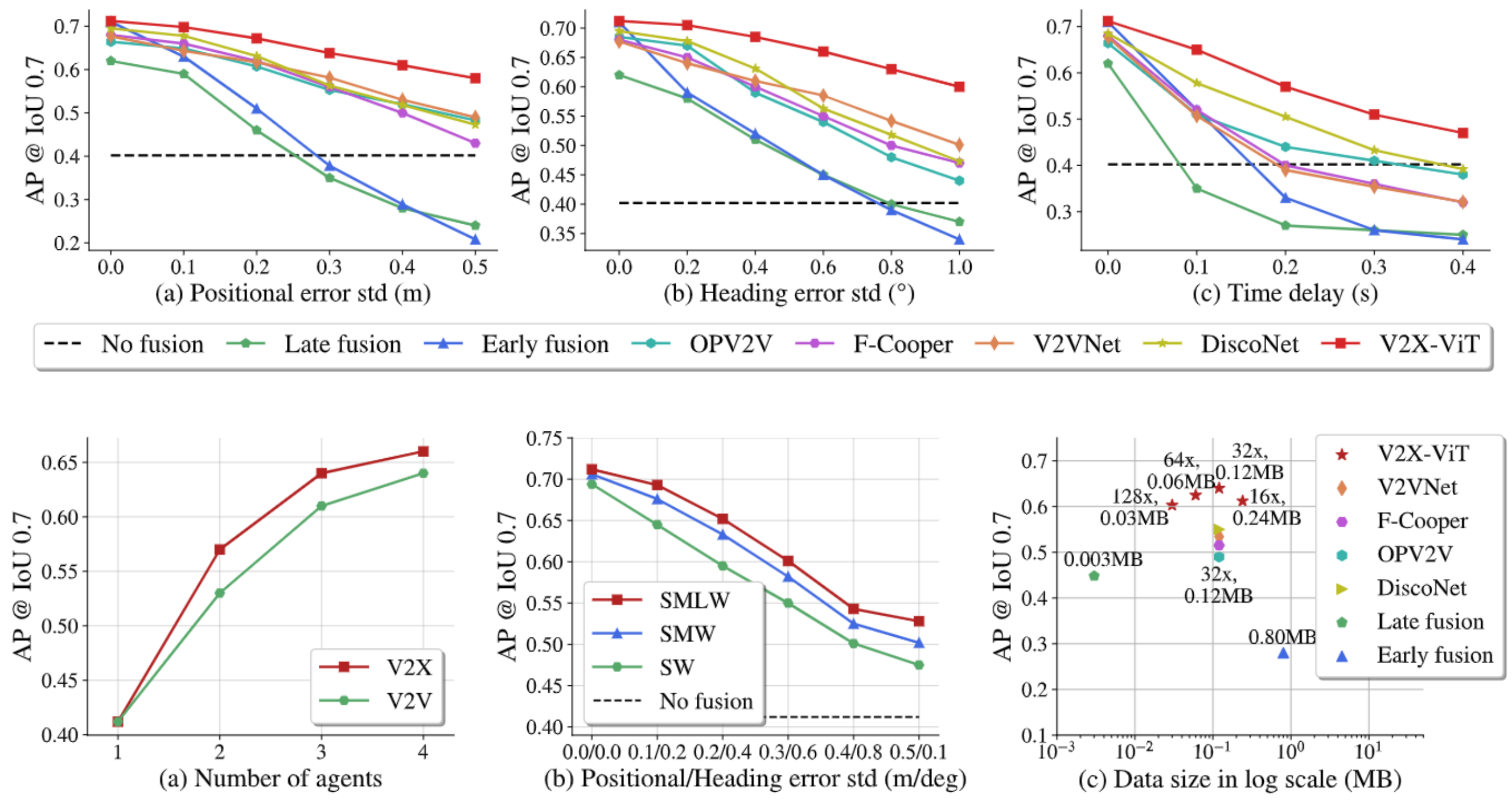


## Benchmark results

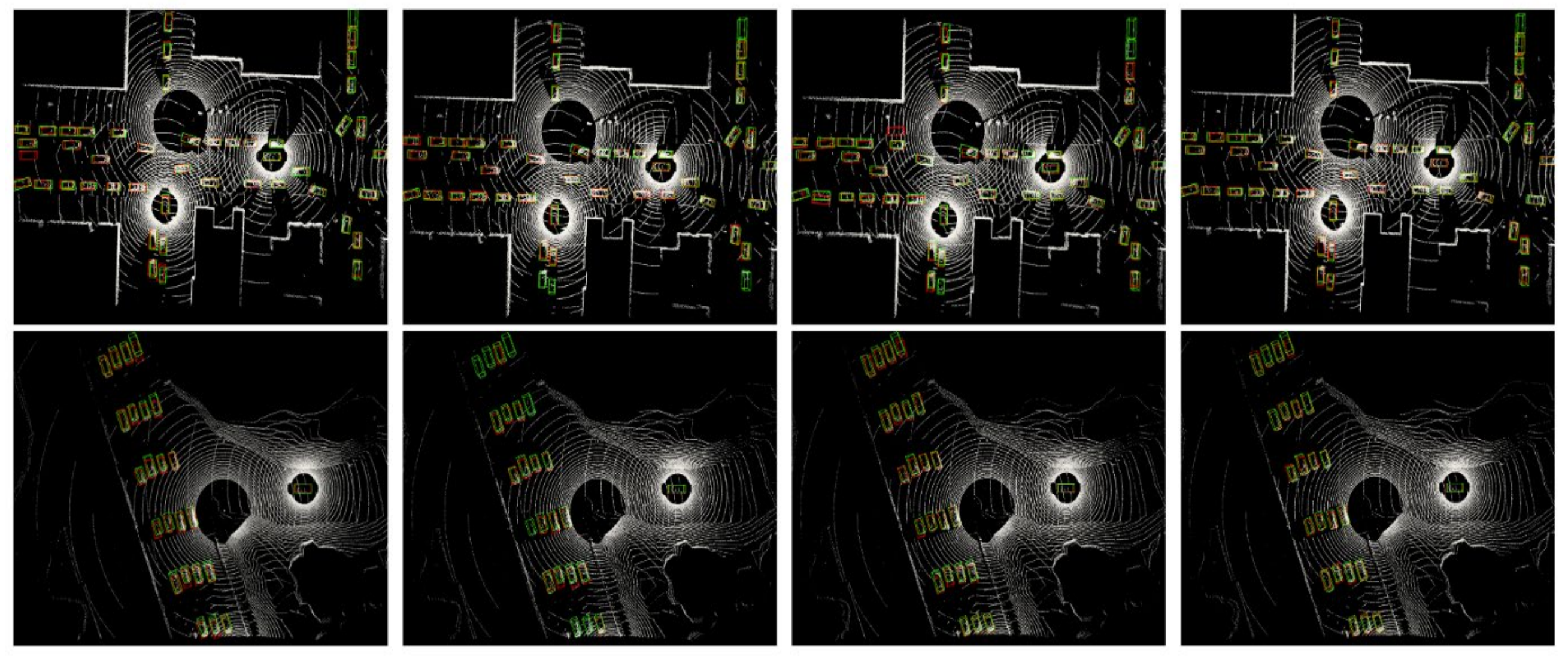
| Models | Perfect AP0.5 | Perfect AP0.7 | Noisy AP0.5 | Noisy AP0.7 |
|---|---|---|---|---|
| No Fusion | 0.606 | 0.402 | 0.606 | 0.402 |
| Late Fusion | 0.727 | 0.620 | 0.549 | 0.307 |
| Early Fusion | 0.819 | 0.710 | 0.720 | 0.384 |
| F-Cooper [4] | 0.840 | 0.680 | 0.715 | 0.469 |
| OPV2V [44] | 0.807 | 0.664 | 0.709 | 0.487 |
| V2VNet [39] | 0.845 | 0.677 | 0.791 | 0.493 |
| DiscoNet [21] | 0.844 | 0.695 | 0.798 | 0.541 |
| V2X-ViT (Ours) | **0.882** | **0.712** | **0.836** | **0.614** |

## Ablation study

| MSwin | SpAttn | HMSA | DPE | AP0.5 / AP0.7 |
|---|---|---|---|---|
| | | | | 0.719 / 0.478 |
| ✓ | | | | 0.748 / 0.519 |
| ✓ | ✓ | | | 0.786 / 0.548 |
| ✓ | ✓ | ✓ | | 0.823 / 0.601 |
| ✓ | ✓ | ✓ | ✓ | **0.836 / 0.614** |

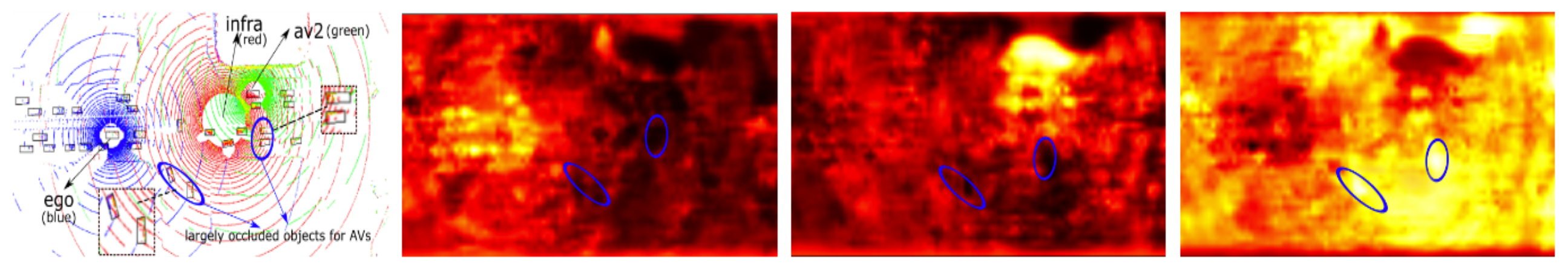


## Detection results



(a) OPV2V [44] (b) V2VNet [39] (c) DiscoNet [21] (d) V2X-ViT (ours)

## Attention map visualization



(a) LiDAR points (better zoom-in) (b) attention weights ego paid to ego (c) attention weights ego paid to av2 (d) attention weights ego paid to infra