AI Model Benchmarking

V2X-ViT

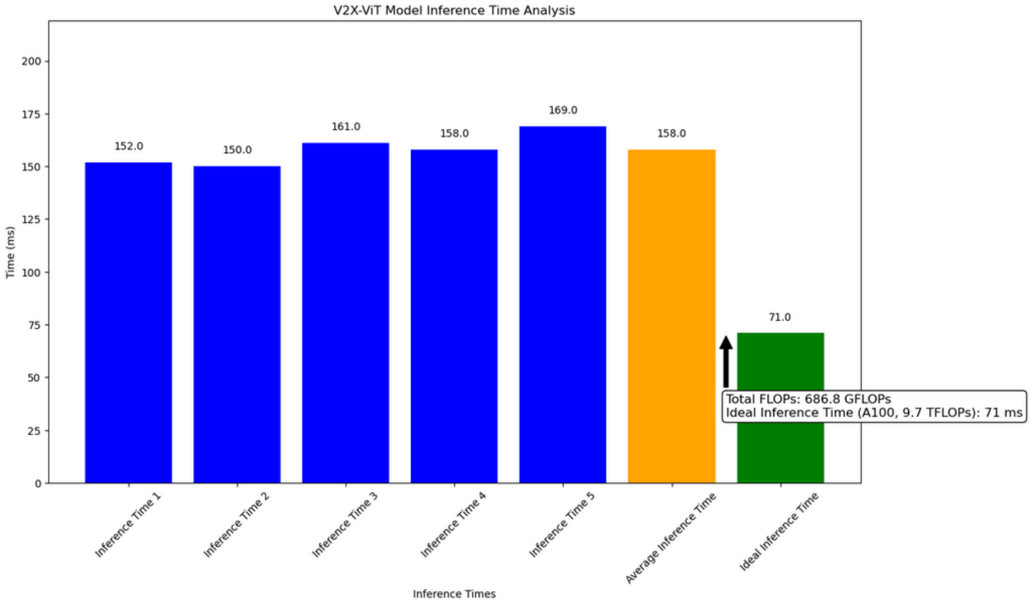
The V2X-ViT model is a vision Transformer designed to enhance the perception performance of autonomous vehicles by leveraging Vehicle-to-Everything (V2X) communication. It consists of alternating layers of heterogeneous multi-agent self-attention and multi-scale window self-attention, allowing it to effectively fuse information from on-road agents such as vehicles and infrastructure. This approach captures inter-agent interaction and per-agent spatial relationships, addressing common V2X challenges like asynchronous information sharing, pose errors, and heterogeneity of V2X components. The model achieves state-of-the-art performance for 3D object detection in noisy environments by integrating these key modules into a unified Transformer architecture.

Benchmark result from the research paper:

- GPU: Nvidia Tesla V100
- Inference time: 57ms for a frame
- Time delay: <400ms
- Complexity O(N)

Benchmark result from my experiment in cluster:

- Dataset: V2V4Real
- GPU: Nvidia Tesla A100
- Inference time: 152ms, 150ms, 161ms, 158ms, 169ms => Average inference time: 158ms
- Total FLOPs: 686.8 GFLOPs
- Ideal inference time (A100 GPU, assume double precision performance, 9.7 TFLOPS) = 686.8 GFLOPs / 9.7 TFLOPS = 71ms



Benchmark Nvidia V100 vs Nvidia A100:

- Result from Lambda Labs: A100 is 60% faster than V100
- Result from Nvidia: A100 is 60-70% faster than V100
- Result from unknow lab: A100 is 30% faster than V100

  => Something wrong with inference time from research paper and my experiment in cluster

Cooperative 3D object detection results from training V2X-ViT from scratch to epoch60

Range	mAP-50	mAP-70
Average	61.5	34.9
Short	81.2	53.4
Middle	49.9	27.8
Long	38.9	15.5

CoBEVT

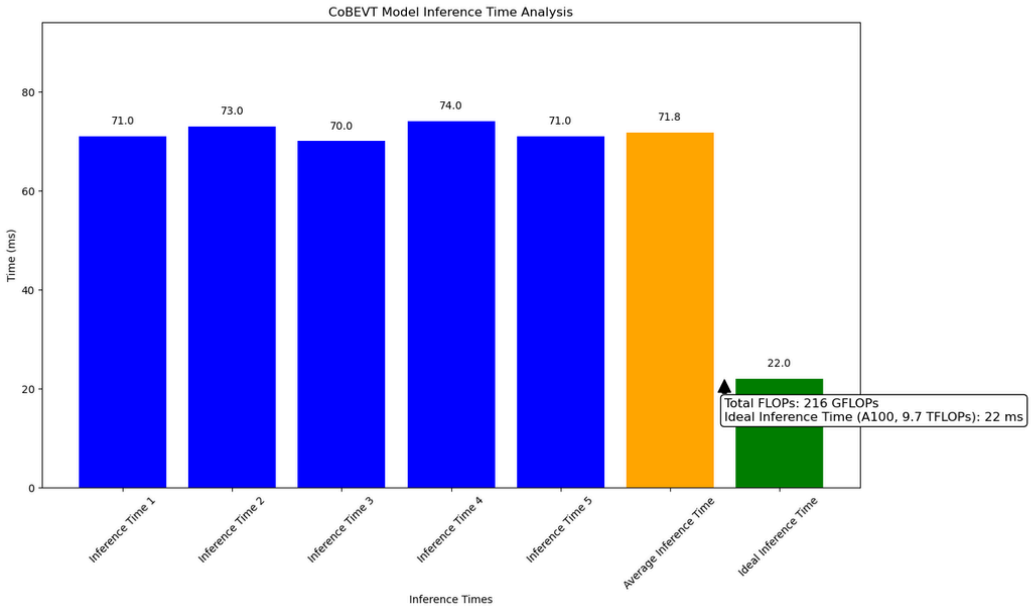
CoBEVT is a novel framework for cooperative bird's eye view (BEV) semantic segmentation in autonomous driving. It addresses limitations of single-agent camera-based systems by leveraging vehicle-to-vehicle (V2V) communication. The framework employs a Transformer architecture with a novel fused axial attention module (FAX) to efficiently fuse features from multi-view and multi-agent data. CoBEVT achieves state-of-the-art performance in cooperative BEV semantic segmentation and is generalizable to other tasks like single-agent multi-camera BEV segmentation and multi-agent LiDAR-based 3D object detection. The model consists of three main components: SinBEVT for single-agent BEV feature computation, feature compression, and sharing; FuseBEVT for multi-agent BEV feature fusion; and a decoder for generating final segmentation output. Extensive experiments demonstrate the effectiveness and efficiency of CoBEVT.

Benchmark result from the research paper:

- GPU: RTX3090
- Inference time: (they said it’s real-time?)
- Complexity: O(N)

Benchmark result from my experiment in cluster:

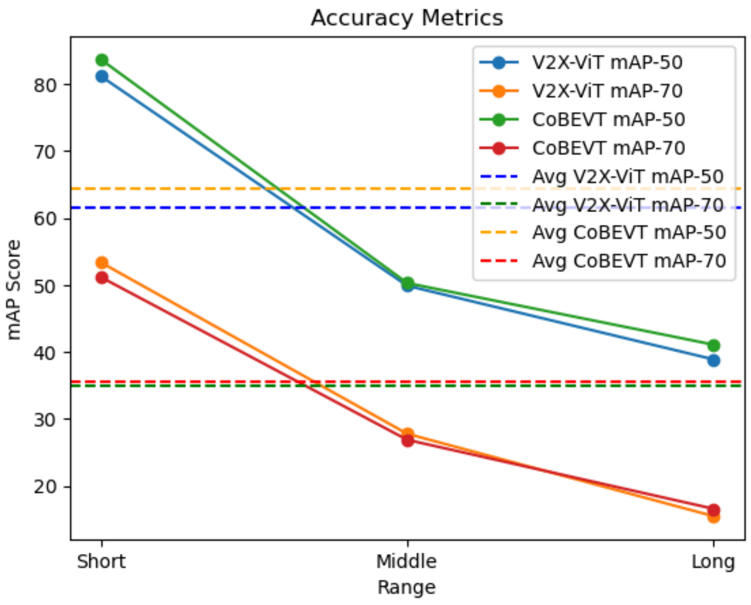
- Dataset: V2V4Real
- GPU: Nvidia Tesla A100
- Inference time: 71ms, 73ms, 70ms, 74ms, 71ms => Average inference time: 72ms
- Total FLOPs: 216 GFLOPs
- Ideal inference time (A100 GPU, assume double precision performance, 9.7 TFLOPS) = 216 GFLOPs / 9.7 TFLOPS = 22ms

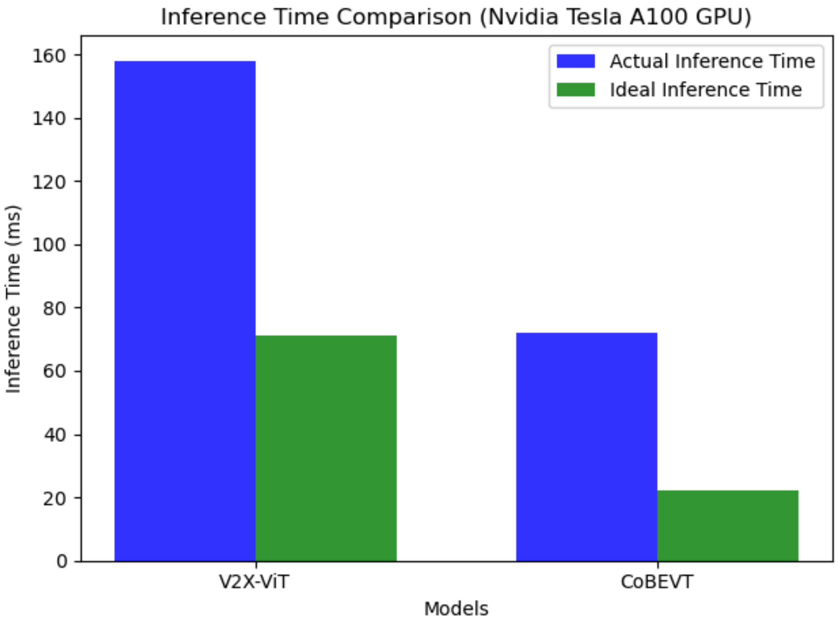


Cooperative 3D object detection results from training CoBEVT from scratch to epoch60

Range	mAP-50	mAP-70
Average	64.5	34.9
Short	83.7	51.2
Middle	50.3	26.9
Long	41.1	16.6

V2X-ViT vs CoBEVT


Accuracy Metrics
V2X-ViT mAP-50
V2X-ViT mAP-70
CoBEVT mAP-50
CoBEVT mAP-70
Avg V2X-ViT mAP-50
Avg V2X-ViT mAP-70
Avg CoBEVT mAP-50
Avg CoBEVT mAP-70
mAP Score
Short
Middle
Long
Range


Inference Time Comparison (Nvidia Tesla A100 GPU)
Actual Inference Time
Ideal Inference Time
Inference Time (ms)
V2X-ViT
CoBEVT
Models