Towards Open-Vocabulary Multimodal 3D Object Detection with Attributes

Xinhao Xiang[1], Kuan-Chuan Peng[2], Suhas Lohit[2], Michael J. Jones[2], Jiawei Zhang[1]

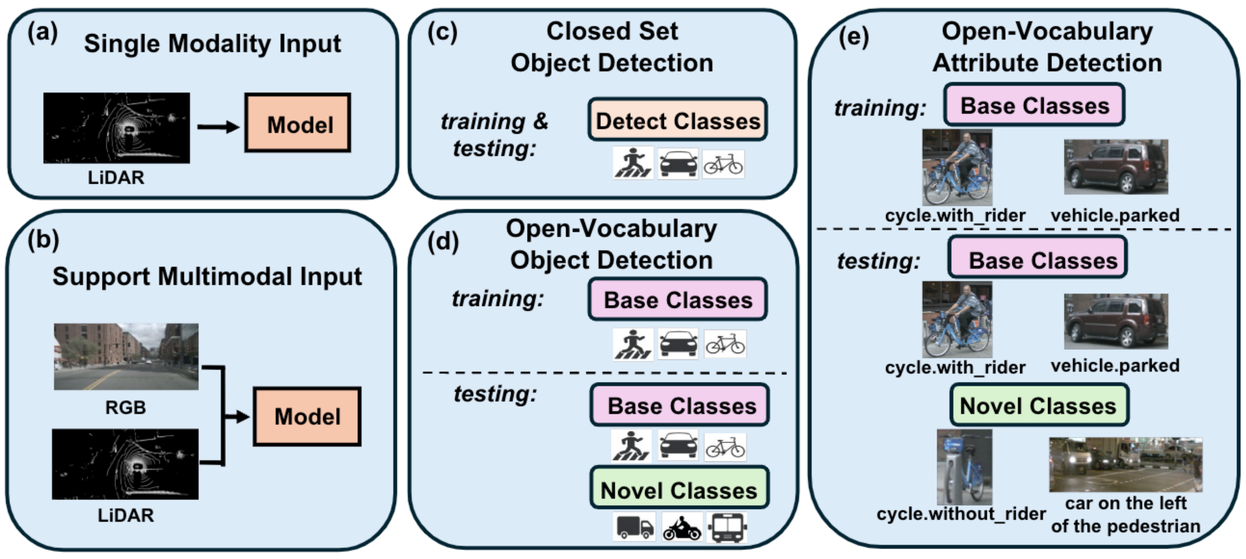
[1] UC DAVIS UNIVERSITY OF CALIFORNIA

[2] MITSUBISHI ELECTRIC Changes for the Better

MITSUBISHI ELECTRIC RESEARCH LABORATORIES

BMVC 2025

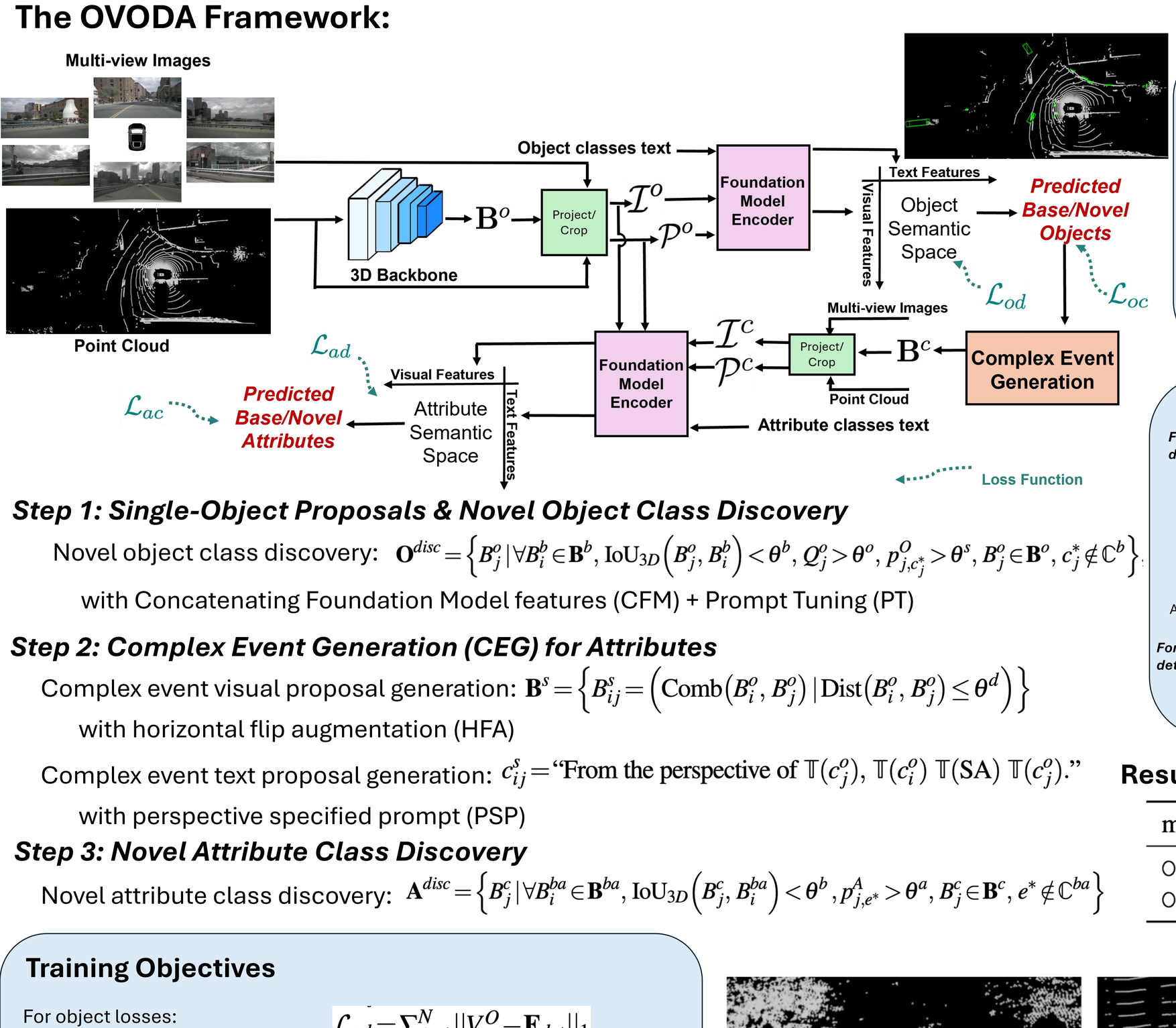
Motivation:



Contribution Summary:

Object/Attribute Detection Conditions	Method Categories							
	C_0	C_1	C_2	C_3	C_4	C_5	C_6	OVODA (ours)
support 3D object detection	✓	✗	✓	✓	✓	✓	✗	✓
can detect attributes	✗	✗	✗	✗	✗	✓	✗	✓
support OV object detection	✗	✓	✓	✓	✓	✗	✗	✓
support multi-modal input	✗	✗	✓	✗	✓	✗	✗	✓
need **no** novel class anchor size	✗	✓	✗	✓	✓	✓	✗	✓
can detect complex events	✗	✗	✗	✗	✗	✗	✓	✓

The OVODA Framework:



Step 1: Single-Object Proposals & Novel Object Class Discovery

Novel object class discovery: $\mathbf{O}^{disc}=\left\{B_j^o \mid \forall B_i^b \in \mathbf{B}^b, \mathrm{IoU}_{3D}\left(B_j^o, B_i^b\right)<\theta^b, Q_j^o>\theta^o, p_{j,c_j^*}^O>\theta^s, B_j^o \in \mathbf{B}^o, c_j^* \notin \mathbb{C}^b\right\}$

with Concatenating Foundation Model features (CFM) + Prompt Tuning (PT)

Step 2: Complex Event Generation (CEG) for Attributes

Complex event visual proposal generation: $\mathbf{B}^s=\left\{B_{ij}^s=\left(\mathrm{Comb}\left(B_i^o, B_j^o\right) \mid \mathrm{Dist}\left(B_i^o, B_j^o\right) \leq \theta^d\right)\right\}$

with horizontal flip augmentation (HFA)

Complex event text proposal generation: $c_{ij}^s=$ "From the perspective of $\mathbb{T}(c_j^o)$, $\mathbb{T}(c_i^o)$ $\mathbb{T}(\mathrm{SA})$ $\mathbb{T}(c_j^o)$."

with perspective specified prompt (PSP)

Step 3: Novel Attribute Class Discovery

Novel attribute class discovery: $\mathbf{A}^{disc}=\left\{B_j^c \mid \forall B_i^{ba} \in \mathbf{B}^{ba}, \mathrm{IoU}_{3D}\left(B_j^c, B_i^{ba}\right)<\theta^b, p_{j,e^*}^A>\theta^a, B_j^c \in \mathbf{B}^c, e^* \notin \mathbb{C}^{ba}\right\}$

OVAD: The Attribute Benchmark

dataset	vocabulary set for attribute detection
OVAD	with rider, without rider, moving, standing, sitting lying down, parked, moving, stopped, in front of, behind, on the left of, on the right of

- Contain 84,384 annotations across 10 classes in total
- To create spatial attribute annotations:

$$\mathbf{B}^{\mathrm{OVAD}}=\left\{B_{ij}^{\mathrm{OVAD}}=(\mathrm{Comb}(B_i, B_j) \mid \mathrm{Dist}(B_i, B_j) \leq 15\mathrm{m})\right\}$$

Experimental Setup

For object detection:

	dataset setting	base object class	novel object class
nuScenes	N_{b6n4}	Car, Construction vehicles, Trailer, Barrier, Bicycle, Pedestrian	Truck, Bus, Motorcycle, Traffic cone
	N_{b3n7}	Car, Bicycle, Pedestrian	Construction vehicles, Trailer, Barrier, Truck, Bus, Motorcycle, Traffic cone
	N_{b0n10}	∅	Car, Construction vehicles, Trailer, Barrier, Bicycle, Pedestrian Truck, Bus, Motorcycle, Traffic cone
Argoverse 2	A_{b4n4}	Regular Vehicle, Trailer, Bicycle, Pedestrian	Truck, Bus, Motorcycle, Construction cone

For attribute detection:

	dataset setting	base attribute class	novel attribute class
OVAD	*OVAD*	with rider, sitting lying down, parked, in front of, behind	without rider, standing, moving, on the left of

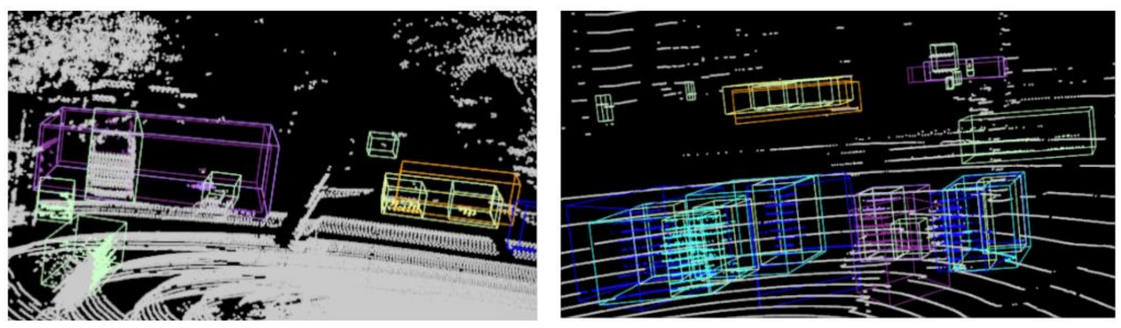
Results: Complex Event (Attribute) Detection

method	CFM	prompt tuning	mAP	AP_{N20}
OVODA	✗	✗	17.24	8.23
OVODA	✓	✓	**27.43**	**12.34**

Training Objectives

For object losses:
- 3D-2D feature distance align: $\mathcal{L}_{od}=\sum_{j=1}^{N}\left\|V_j^O-\mathbf{F}_{det}\right\|_1$
- 3D-Text classification contrastive: $\mathcal{L}_{oc}=\sum_{j=1}^{N} f\left(B_j^{disc}, \mathbf{B}^b\right) \cdot CE\left(\mathbf{P}_j^{disc}, h_j^O\right)$

For attribute losses:
- 3D-2D feature distance align: $\mathcal{L}_{ad}=\sum_{j=1}^{N}\left\|V_j^A-\mathbf{F}_c\right\|_1$
- 3D-Text classification contrastive: $\mathcal{L}_{ac}=\sum_{j=1}^{N} g\left(A_j^{disc}, \mathbf{B}^{ba}\right) \cdot CE\left(\mathbf{P}_j^{disc,a}, h_j^A\right)$



Ablations:

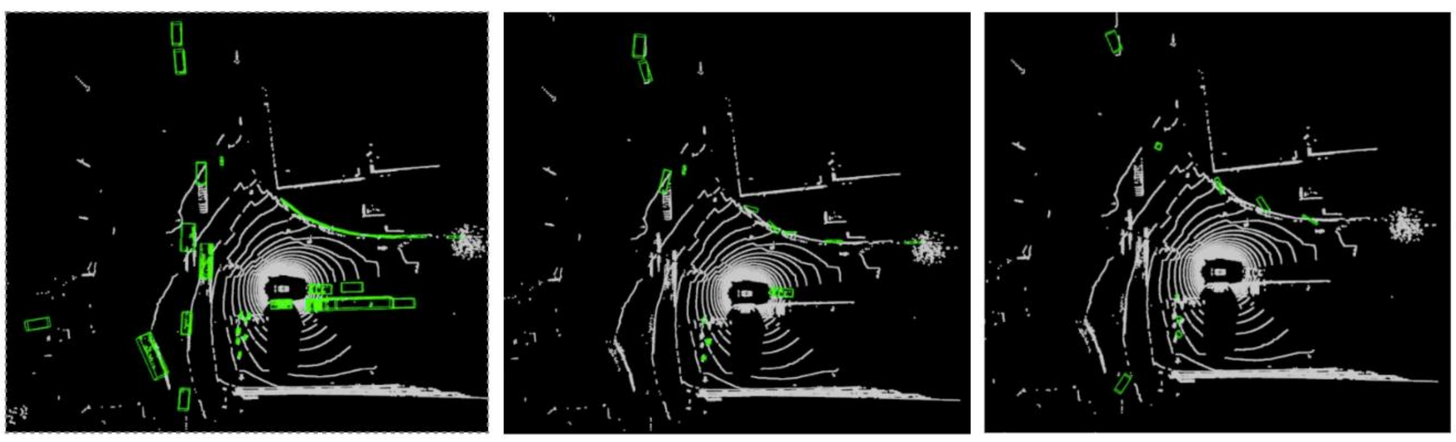
method	CFM	PT	mAP	NDS	AP_{N20}	SR (%) (AD only)	SR (%) (AD & OD)
OVODA	✗	✗	30.24	31.54	14.24	16.35	4.23
	✓	✗	31.34	31.63	14.42	22.74	5.56
	✓	✓	**32.25**	**31.85**	**14.72**	**25.90**	**6.77**

method	foundation model	mAP	NDS	AP_{N20}	SR (%) (AD only)	SR (%) (for both AD & OD)
OVODA	CLIP [33]	21.93	21.54	11.45	16.20	3.22
	CogVLM [40]	25.34	26.81	12.84	19.84	5.39
	OneLLM [14]	**32.25**	**31.85**	**14.72**	**25.90**	**6.77**

method	PSP	HFA	FM	mAP	NDS	AP_{N20}	SR (%) (AD only)	SR (%) (for both AD & OD)
OVODA	✗	✗	CLIP	21.93	21.54	11.45	16.20	3.22
	✓	✗	CLIP	22.34	24.35	11.94	16.83	3.49
	✗	✓	CLIP	22.46	23.43	12.32	17.43	4.03
	✓	✓	CLIP	24.37	25.83	13.02	19.42	4.92
	✓	✓	OneLLM	**32.25**	**31.85**	**14.72**	**25.90**	**6.77**

Results: Open-Vocabulary 3D Object Detection

method	CFM	prompt tuning	need **no** predefined anchor size for each novel class?	N_{b6n4}			N_{b3n7}			N_{b0n10}		
				mAP	NDS	AP_{N20}	mAP	NDS	AP_{N20}	mAP	NDS	AP_{N20}
Find n′ Propagate [8]	✗	✗	✗	44.95	47.87	33.65	37.38	40.28	18.46	N/A	N/A	16.72
CoDAv2 [5]	✗	✗	✓	27.35	29.48	12.63	18.73	20.14	8.74	4.32	5.82	1.37
OVODA	✗	✗	✓	30.24	31.54	14.24	20.46	21.83	9.31	4.57	6.96	2.16
OVODA	✓	✓	✓	**32.25**	**31.85**	**14.72**	**21.03**	**22.14**	**10.23**	**4.70**	**7.05**	**2.39**



Ground Truth | OVODA's Prediction | CoDAv2's Prediction