OpenBox: Annotate Any Bounding Boxes in 3D

In-Jae Lee

Moongyeom Kim

Kwonyoung Ryu

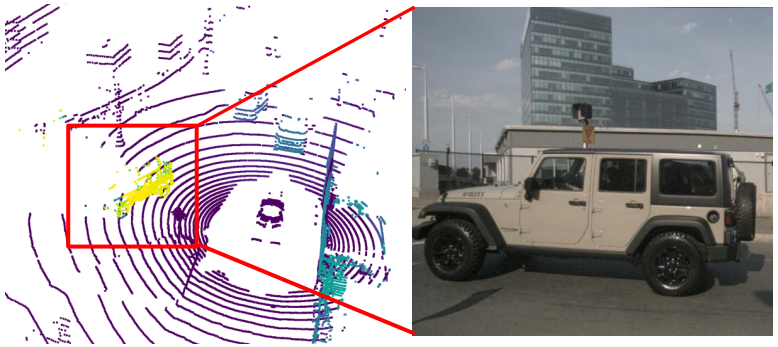
Pierre Musacchio

Jaesik Park

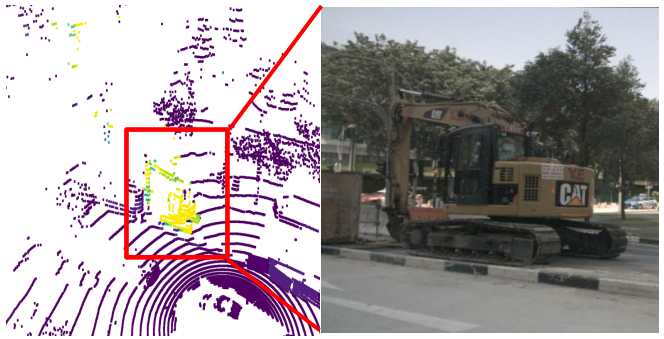
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Open-vocabulary 3D Object Detection



Text query: *Jeep*



Text query: Excavator

Text query: Crowd

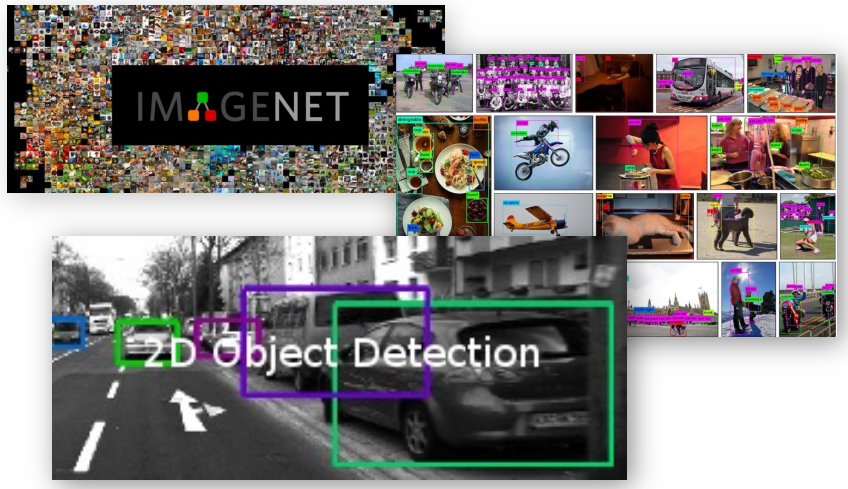
Open-vocabulary detection recognizes **novel classes** in previously **unseen** domains.

- Human Annotation for the Dataset

Human annotation is *costly* and ***labor-intensive***.

Motivation

2D Vision Tasks



Various datasets & labels for tasks

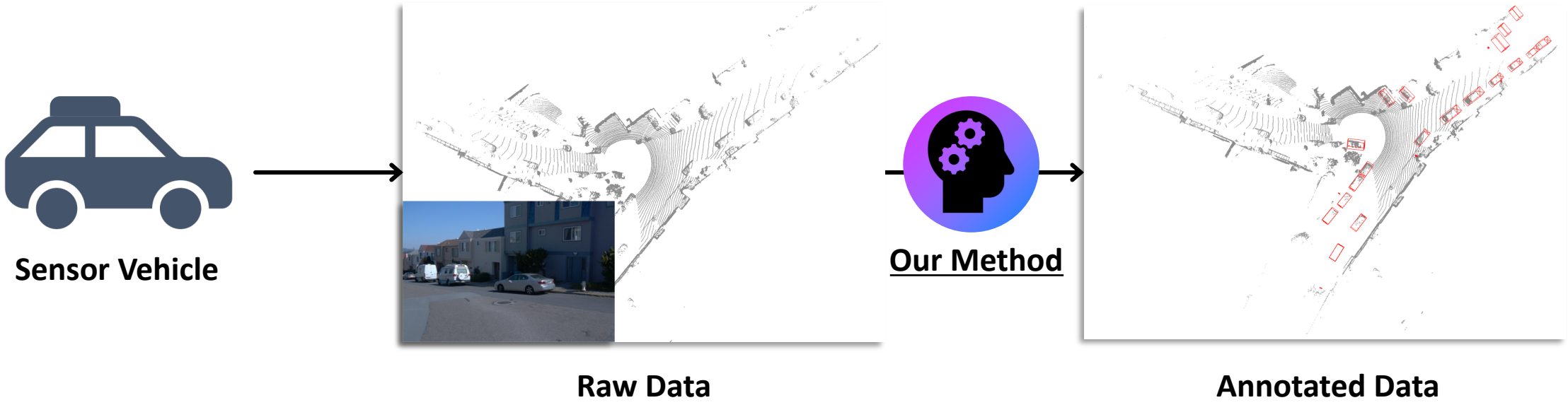
3D Perception Tasks

Limited datasets & labels due to costly annotation

Research Goal

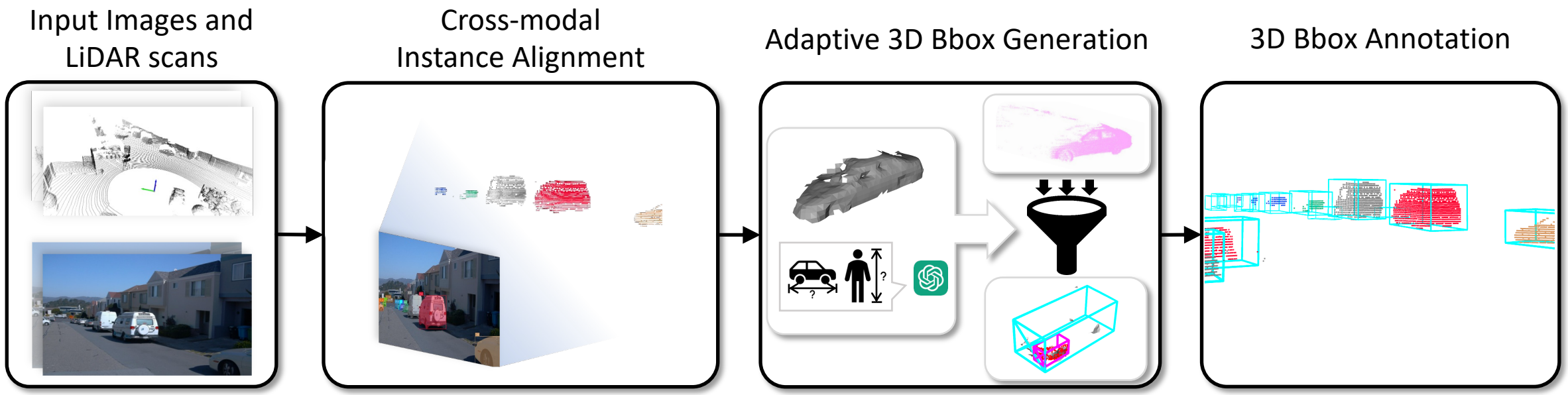
Automatic Annotation *for 3D Object Detection*

Leveraging 2D ***Vision Foundation Model***



Method

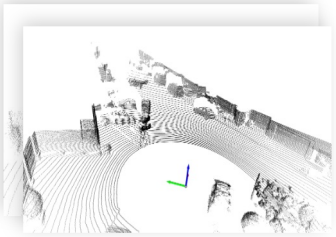
- Overall Pipeline



- Contribution
  1. Only time synchronized ego poses, images and LiDAR point clouds are needed.
  2. High quality point cloud refinement & Adaptive 3D bounding box Generation.
  3. SOTA performance autonomous driving dataset (Waymo, nuScenes, Lyft).

Method (Stage 1: Cross-modal Instance Alignment)

- Instance-level Feature Extraction



Multi-frame LiDAR Point Clouds

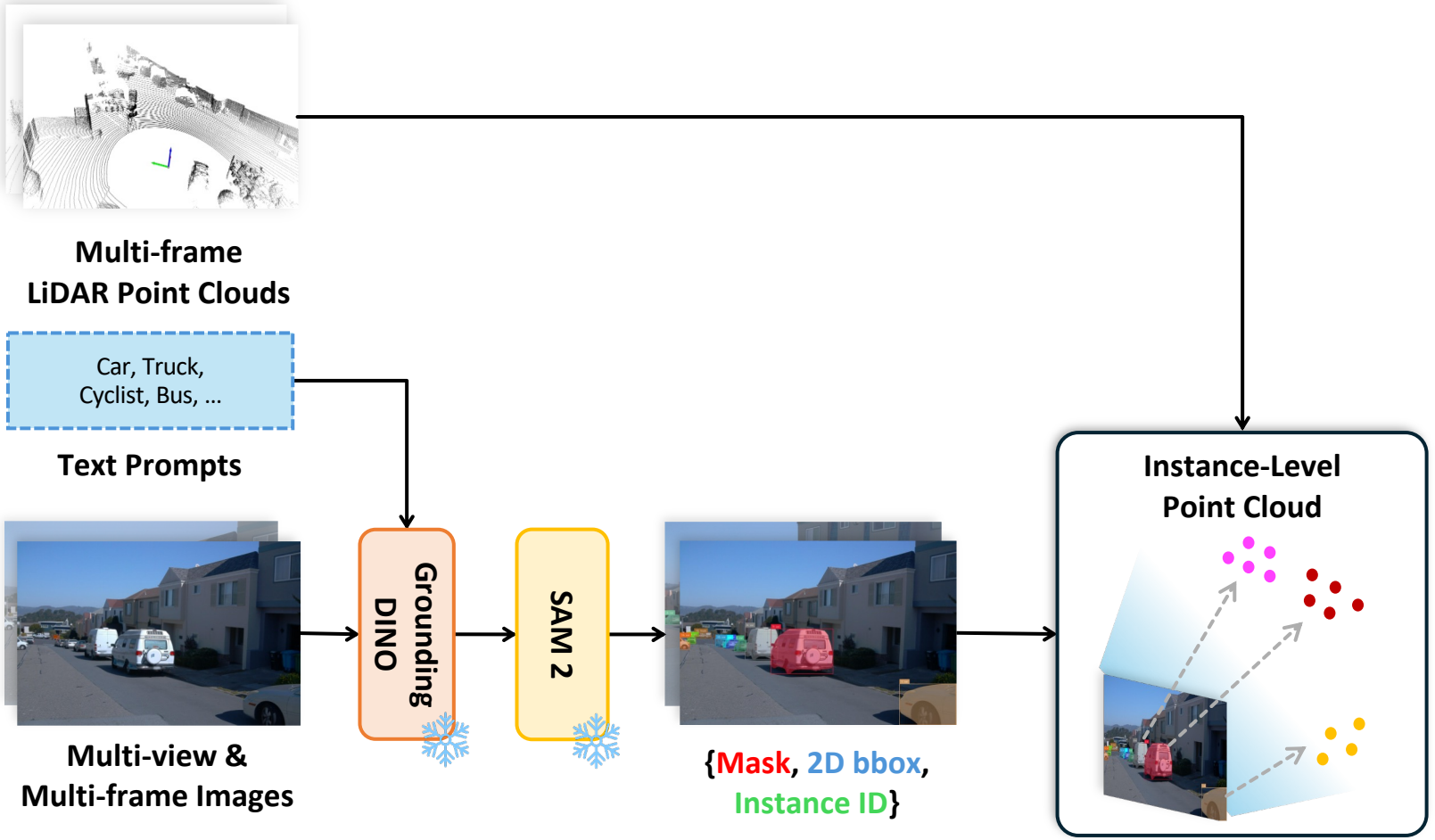
Car, Truck, Cyclist, Bus, ...

Text Prompts

Multi-view & Multi-frame Images

Method (Stage 1: Cross-modal Instance Alignment)

- Instance-level Feature Extraction



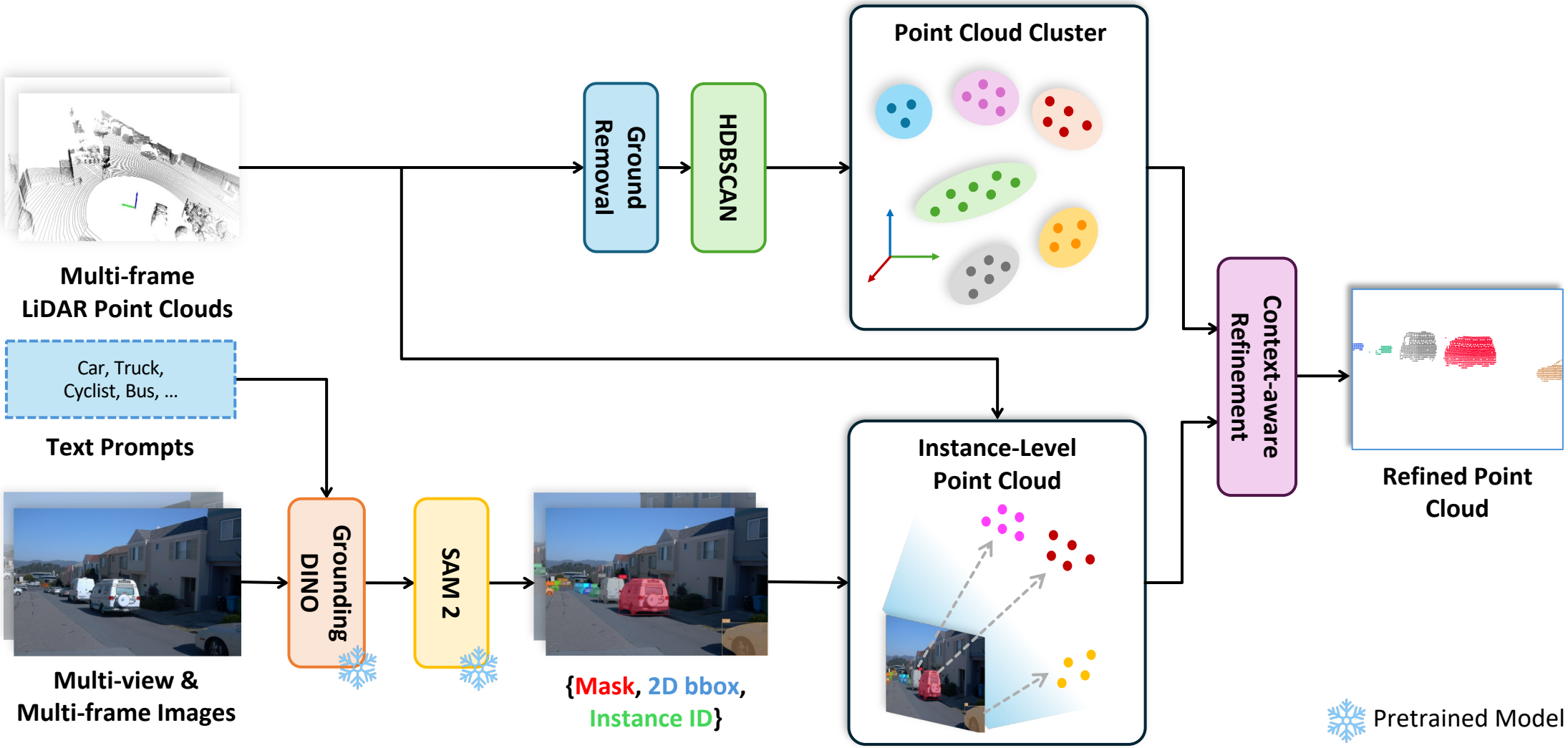
Method (Stage 1: Cross-modal Instance Alignment)

- Context-aware Refinement

(a) Reference Image (I) (b) Point Cloud Cluster ($\mathcal{R}$)

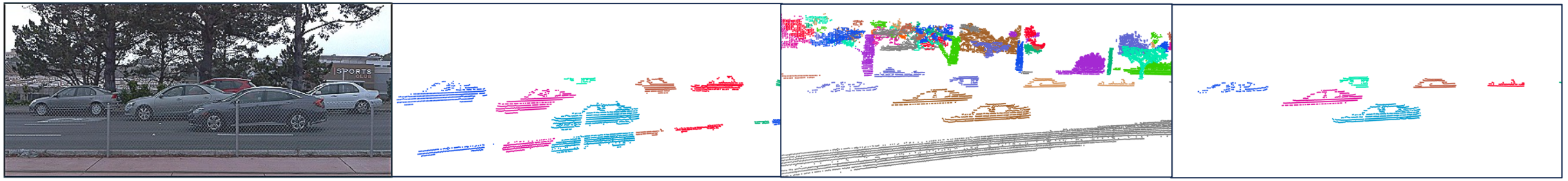
Method (Stage 1: Cross-modal Instance Alignment)

- Context-aware Refinement



Method (Stage 1: Cross-modal Instance Alignment)

- Context-aware Refinement



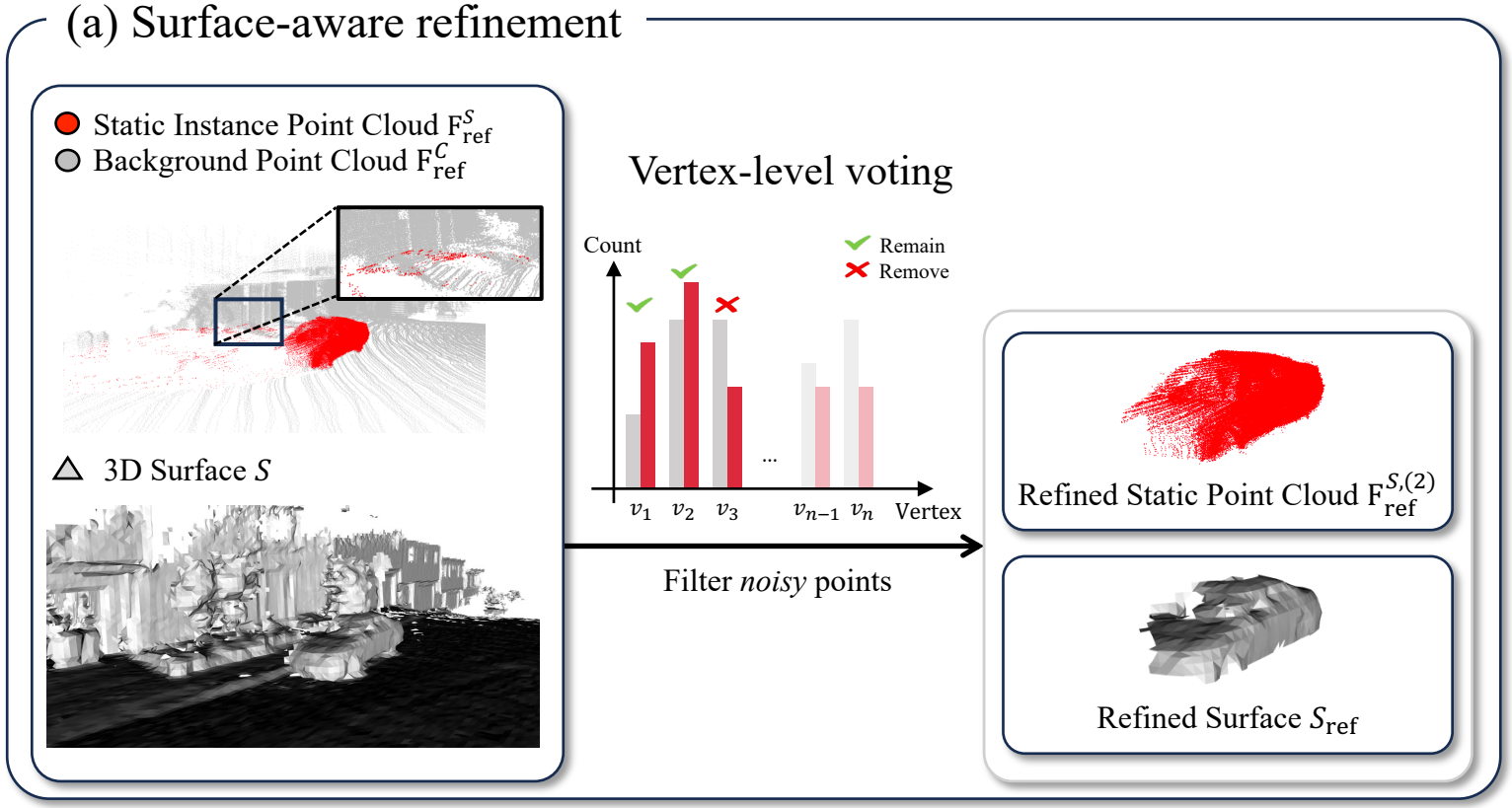
(a) Reference Image (I) (b) Point Cloud Cluster ($\mathcal{R}$) (c) Instance-level Point Cloud ($\mathcal{F}$) (d) Refined Point Cloud ($\mathcal{F}_{\text{ref}}$)

$$\frac{|\{p \in \mathcal{R}_k \mid \text{dist}(p, \mathcal{F}_i) < \delta\}|}{|\mathcal{R}_k|} > \alpha, \quad \frac{|\{p \in \mathcal{R}_k \mid \text{dist}(p, \mathcal{F}_i) < \delta\}|}{|\mathcal{F}_i|} > \beta \quad \Rightarrow \quad \mathcal{R}_k \leftarrow i, \qquad (2)$$

where $|\cdot|$ denotes the cardinality of a set, and $\text{dist}(p, \mathcal{F}_i) < \delta$ holds if and only if there exists $f \in \mathcal{F}_i$ such that $\|p - f\|_2 < \delta$.

Method (Stage 2: Adaptive 3D Bounding Box Generation)

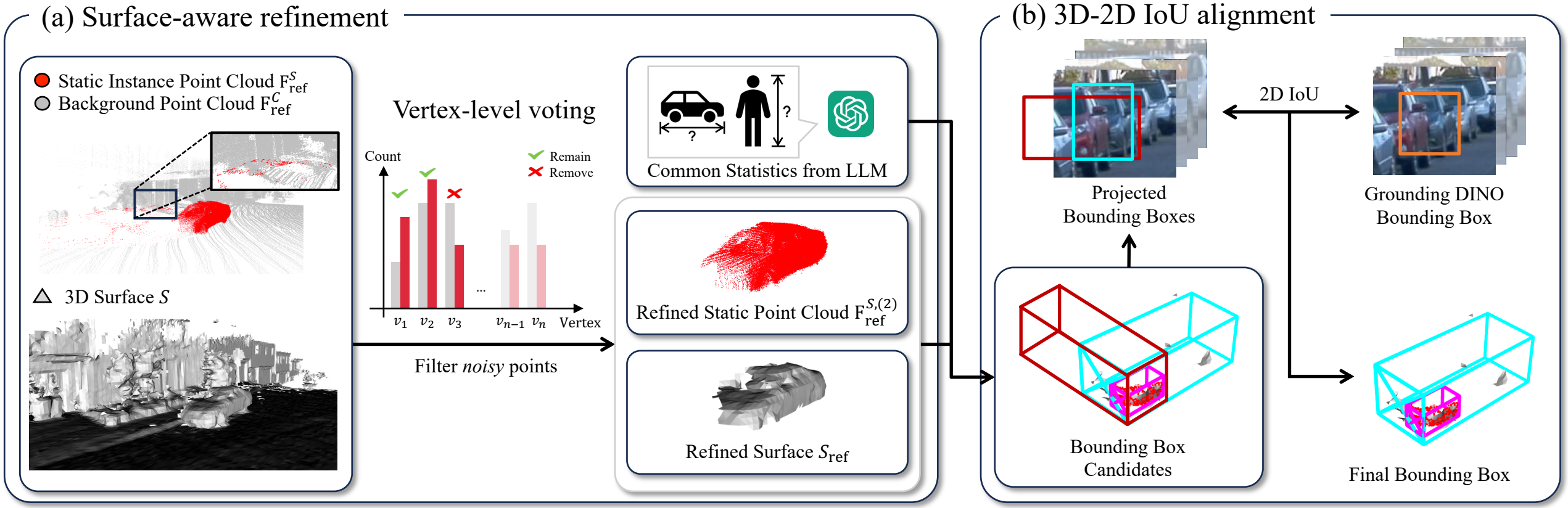
- Rigid Static Object



(a) Instance-level mesh & point cloud refinement via **vertex level voting.**

Method (Stage 2: Adaptive 3D Bounding Box Generation)

- Rigid Static Object



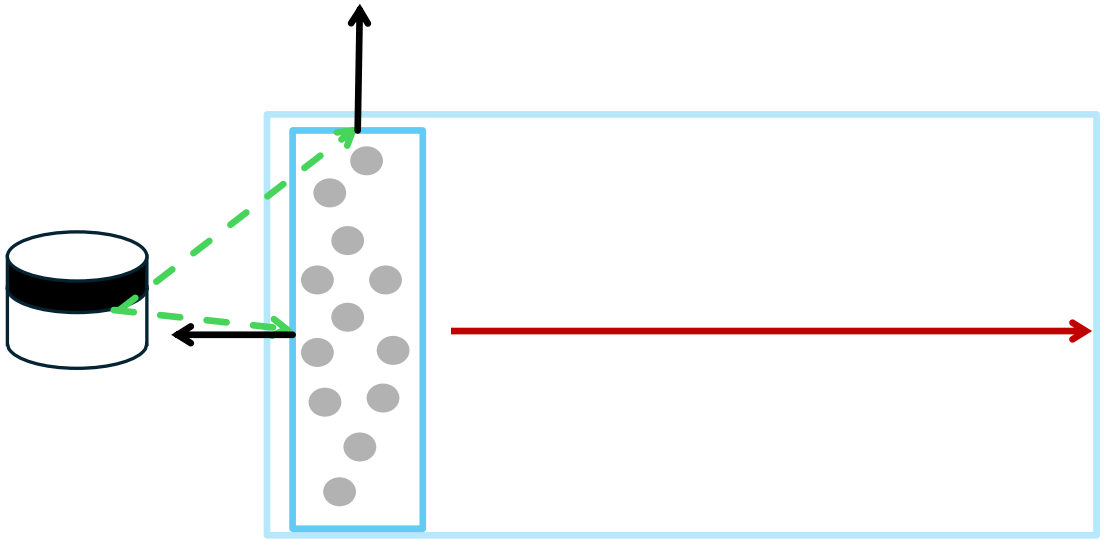
(a) Instance-level mesh & point cloud refinement via **vertex level voting.**

(b) Search optimal bbox by **calculating IoU** between projected 3D boxes and Grounding DINO box.

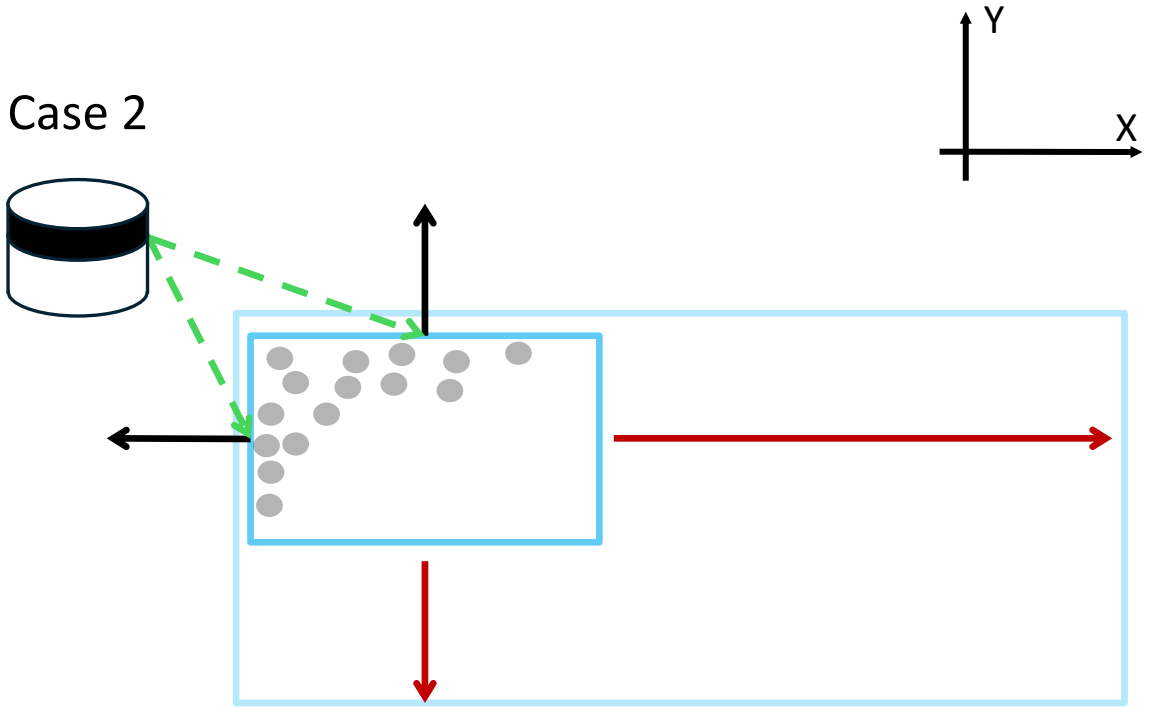
Method (Stage 2: Adaptive 3D Bounding Box Generation)

- Rigid Dynamic Object

Case 1



Case 2



Initial box
Refined box
LiDAR ray
Box normal
Box extension

Experiments

- 3D Object Detection Results on Waymo dataset
  - Setting 1 (Train: Auto annotation → Test: Human annotation)

Method	Modality	Vehicle $IoU_{0.5}$ / $IoU_{0.7}$	Pedestrian $IoU_{0.3}$ / $IoU_{0.5}$	Cyclist $IoU_{0.3}$ / $IoU_{0.5}$
CPD* [40]	LiDAR	30.30 / 20.90	14.28 / 11.22	3.47 / **3.08**
OpenBox* (Ours)	LiDAR + Camera	**70.49 / 32.41**	**57.95 / 17.11**	**20.81** / 2.15
DBSCAN [7]	LiDAR	2.32 / 0.29	0.51 / 0.00	0.28 / 0.03
MODEST [43]	LiDAR	18.51 / 6.46	11.83 / 0.17	1.47 / 1.14
OYSTER [45]	LiDAR	30.48 / 14.66	4.33 / 0.18	1.27 / 0.33
CPD [40]	LiDAR	57.79 / 37.40	21.91 / 16.31	5.83 / 5.06
OpenBox† (Ours)	LiDAR + Camera	**66.89** / 39.14	**55.71 / 37.82**	**21.00 / 7.08**
OpenBox‡ (Ours)	LiDAR + Camera	59.09 / **40.68**	39.09 / 28.16	8.27 / 6.23
Human Annotation	-	93.31 / 75.70	87.25 / 77.93	58.84 / 54.88

Experiments

- 3D Object Detection Results on Lyft dataset
  - Setting 1 (Train: Auto annotation → Test: Human annotation)

Method	Modality	0–30m	30–50m	50–80m	0–80m
MODEST-PP ($T = 0$) [43]	LiDAR	46.4 / 45.4	16.5 / 10.8	0.9 / 0.4	21.8 / 18.0
LiSe ($T = 0$) [46]	LiDAR + Camera	54.5 / 54.0	24.2 / 22.8	1.4 / 1.2	29.2 / 27.5
OpenBox (Ours)	LiDAR + Camera	**62.4 / 62.3**	**56.6 / 50.6**	**19.9 / 19.5**	**49.6 / 43.3**
Human Annotation	-	82.8 / 82.6	70.8 / 70.3	50.2 / 49.6	69.5 / 69.1

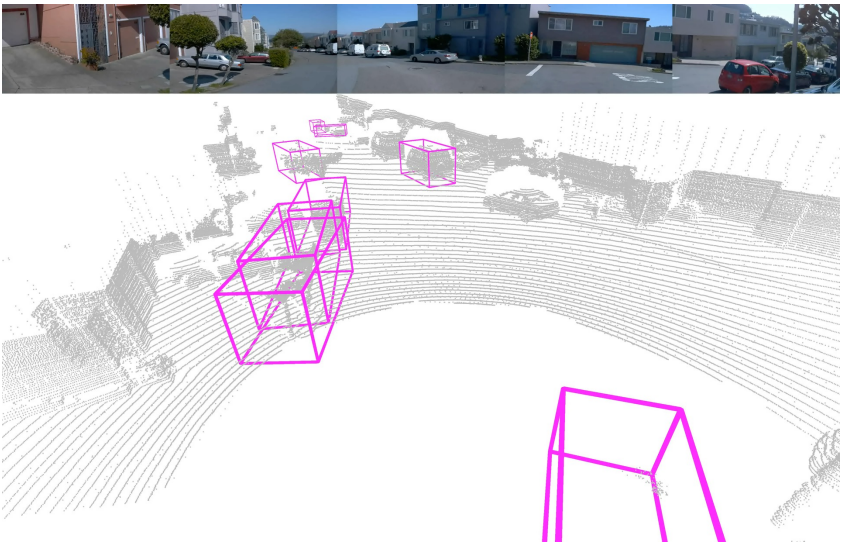
  - Setting 2 (Comparison with auto annotation with human annotation on training set)

Method	0–30m	30–50m	50–80m	0–80m
LiSe [46]	17.47	6.87	1.35	6.31
OpenBox (Ours)	**56.62**	**28.10**	**6.47**	**26.25**

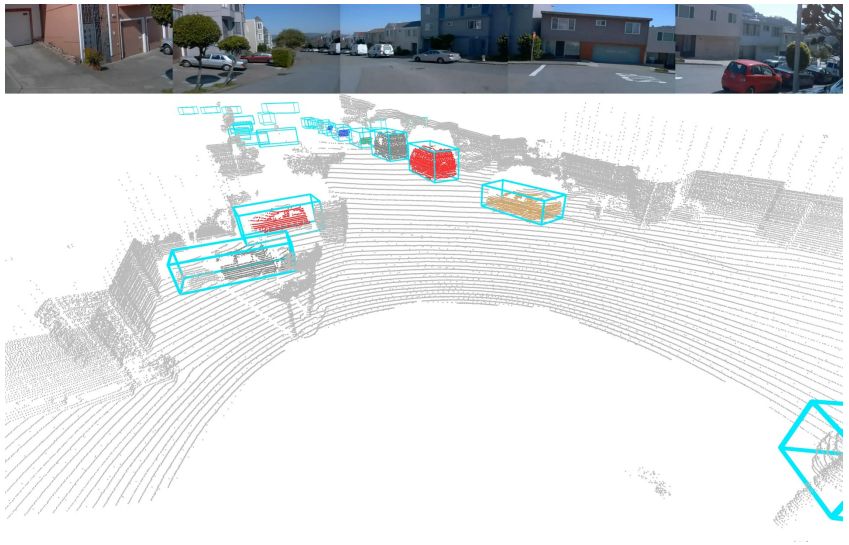
Experiments

- Scenario 1 (Baseline vs. OpenBox (Ours) vs. Human Annotation)

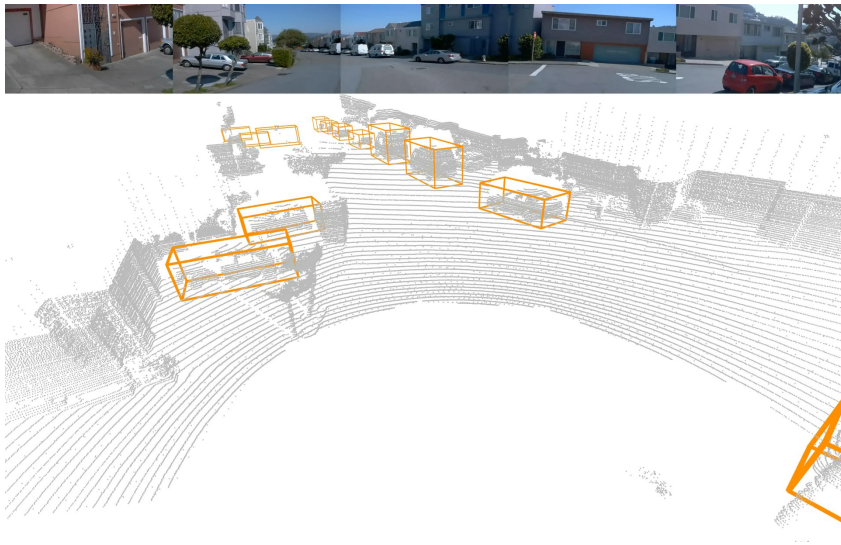
Baseline



OpenBox (Ours)



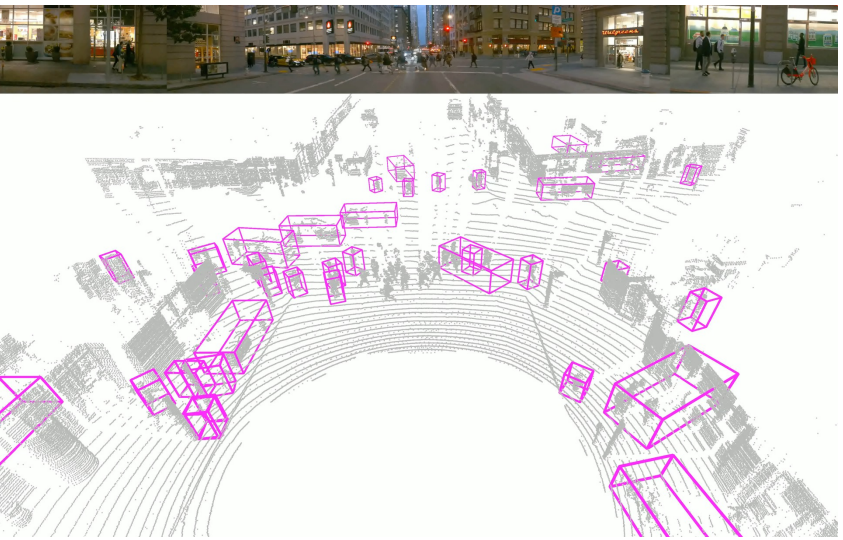
Human Annotation



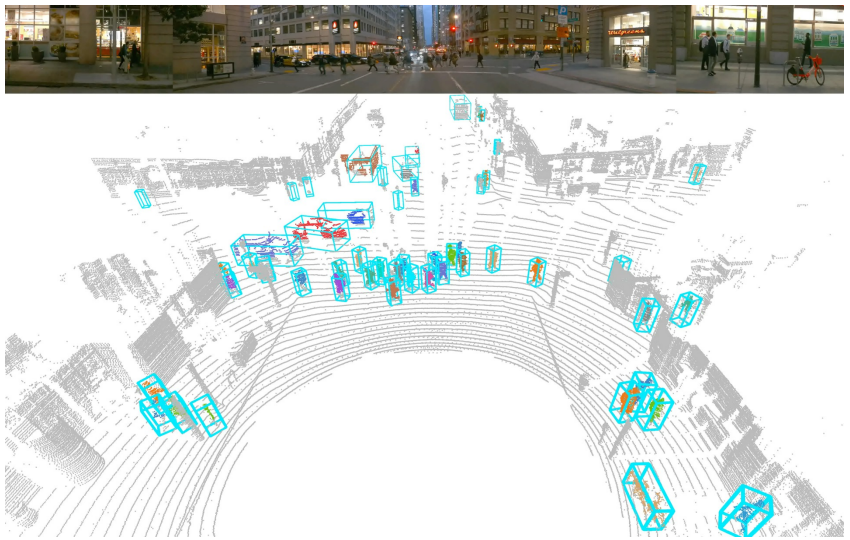
Experiments

- Scenario 1 (Baseline vs. OpenBox (Ours) vs. Human Annotation)

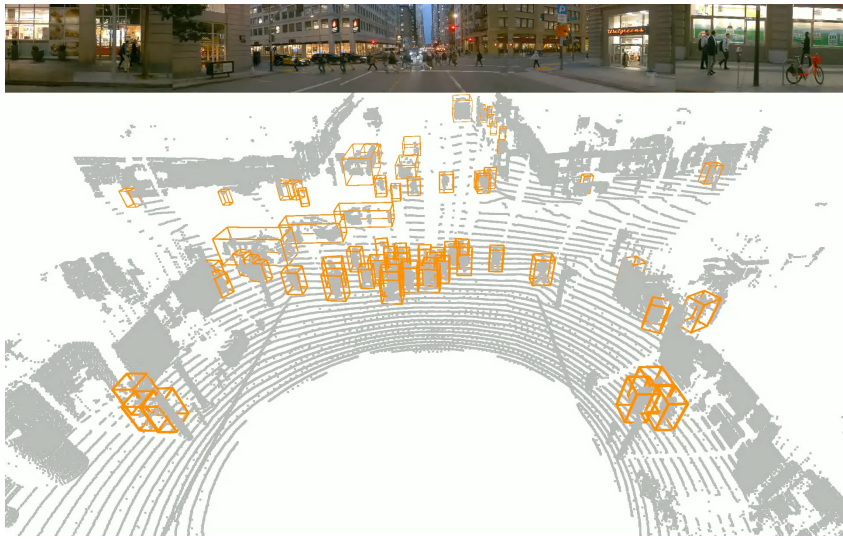
Baseline



OpenBox (Ours)



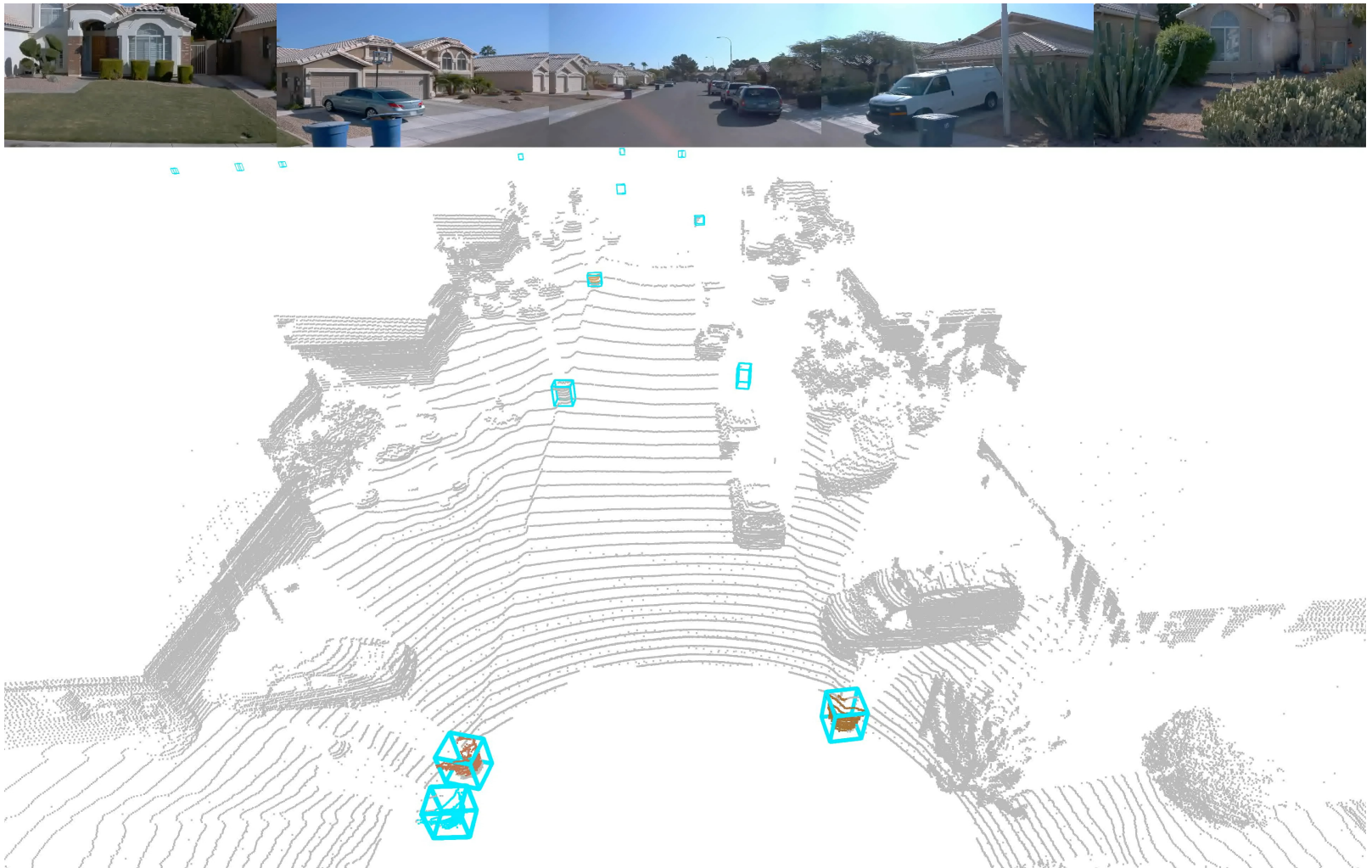
Human Annotation



Experiments

- Scenario 2 (Automatic Open-vocabulary Annotation)

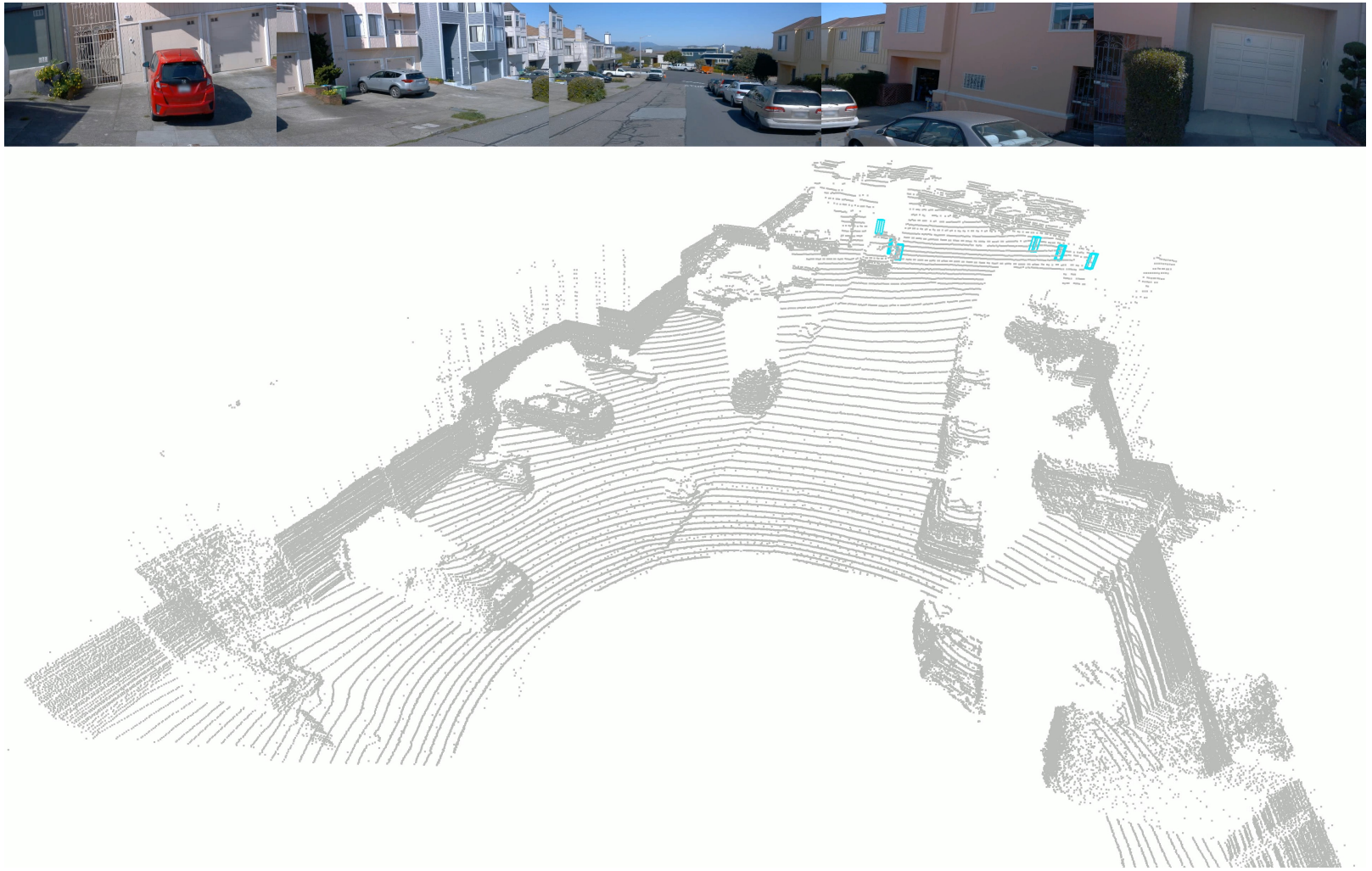
Text Prompt: `Trash bin`



Experiments

- Scenario 2 (Automatic Open-vocabulary Annotation)

Text Prompt: `Traffic Cone`



Thank you for watching!