IEEE IV 2026

Improving 3D Labeling in Self-Driving by Inferring Vehicle Information using Vision Language Models

Steven Chen, Shivesh Khaitan, Nemanja Djuric

Aurora Innovation, Inc.

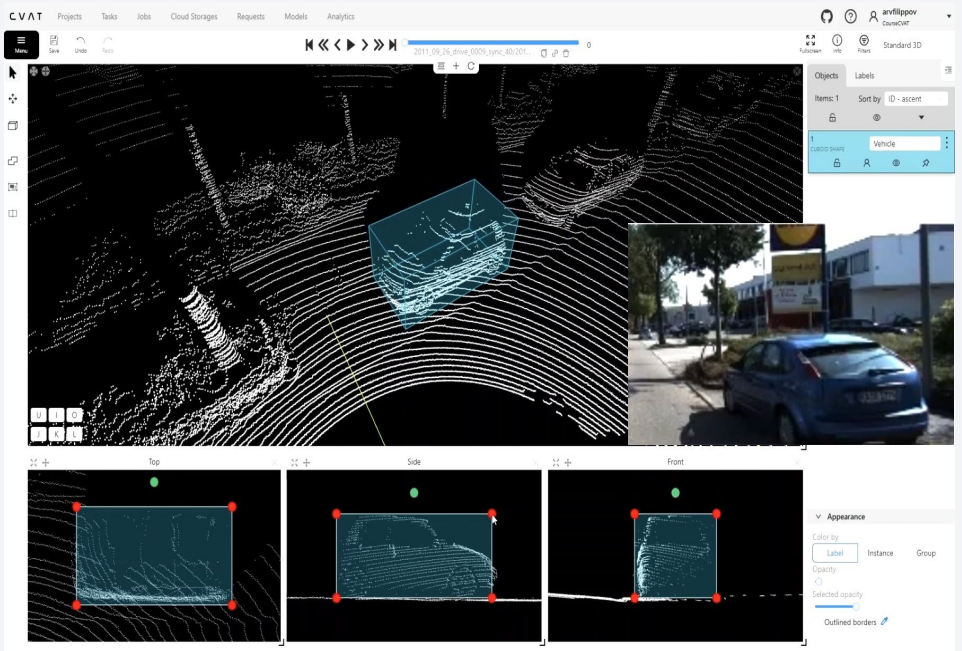
The 3D Labeling Bottleneck

Self-driving systems train on huge amounts of vehicle labels

Labels are hand-drawn by humans: hard, slow, expensive

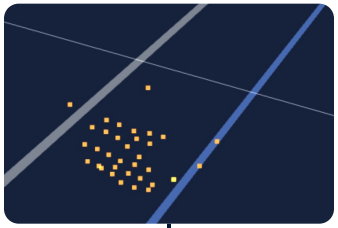
Auto-labeling: generate seed 3D boxes as input to human labeling

- Then, only minor edits needed
- Easier, faster, cheaper

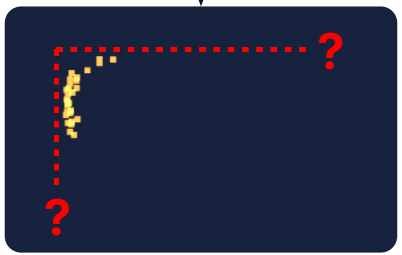


CVAT: an open source 3D labeling tool

Current 3D Auto-labeling



Custom-trained,
Lidar-centric 3D model

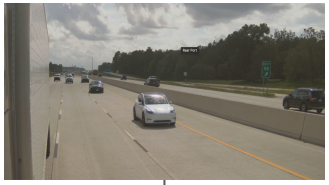


Our Approach

Toyota RAV4 5th Gen (2018-2024)

Factory dimensions: 4.60 × 1.85 × 1.68m

System Pipeline



Vehicle Detection, Viewpoint Selection, Cropping



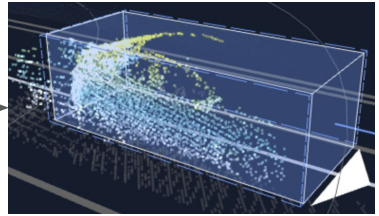
Prompt:
Identify the vehicle's make, model, and generation. Based on that, provide the factory dimensions.

Vision Language Model

Make: **Tesla**
Model: **Model Y**
Generation: **1** (2020-2024)

3D Localized Detection

Combine Detection with VLM Dimensions



Results: Prompt Engineering

If you ask directly for dimensions, VLMs guess

If you force reasoning about make, model, generation (VMMGR), accuracy increases substantially

	VLM Prompt*		
	What are the dimensions?	What **class** (i.e. midsize sedan, full size SUV) → dimensions?	What is the **make, model, generation** → factory dimensions?
Intersection over Union	0.845	0.877	**0.883**
Length Error (m)	0.301	0.281	**0.259**
Width Error (m)	0.116	0.077	**0.077**
Height Error (m)	0.094	0.083	**0.075**

*simplified prompts shown

Results: Predicting Dimensions using various VLMs

	Baseline	Llama 4 Maverick	Pixtral Large	Claude Sonnet 4	Gemini Pro 2.5
Intersection over Union	0.845	0.856	0.864	0.866	**0.883**
Abs Length Error (m)	0.301	0.295	0.284	0.291	**0.259**
Abs Width Error (m)	0.116	0.096	0.091	0.085	**0.077**
Abs Height Error (m)	0.094	0.103	0.095	0.093	**0.075**

Baseline: **Oracle** provided with ground-truth vehicle types

VLMs: Zero-shot inference, uses only vehicle images

Above: Aurora dataset
Also evaluated on Waymo Open Dataset (see the paper for details)

Recognition in Action

Far Away

Chrysler 300
1st generation

Different Configs

Ford Transit Connect
2nd generation

Vintage

Chevy Silverado
1990-1998 (C/K)

Generic Styling

Gemini: Kia Forte 2nd Gen
Other VLMs: similar sedans

Top VLM (Gemini 2.5 Pro) achieved **98.1%** accuracy on make/model/generation recognition

Long Tail: Capturing Modifications

Add to prompt

Is the vehicle **modified**? In what directions (i.e. wider, shorter)

Benefits

Useful for auto-adjusting dimensions

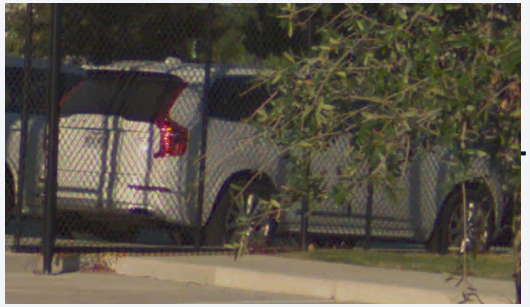
Flags modified vehicles for human review, ML upsampling

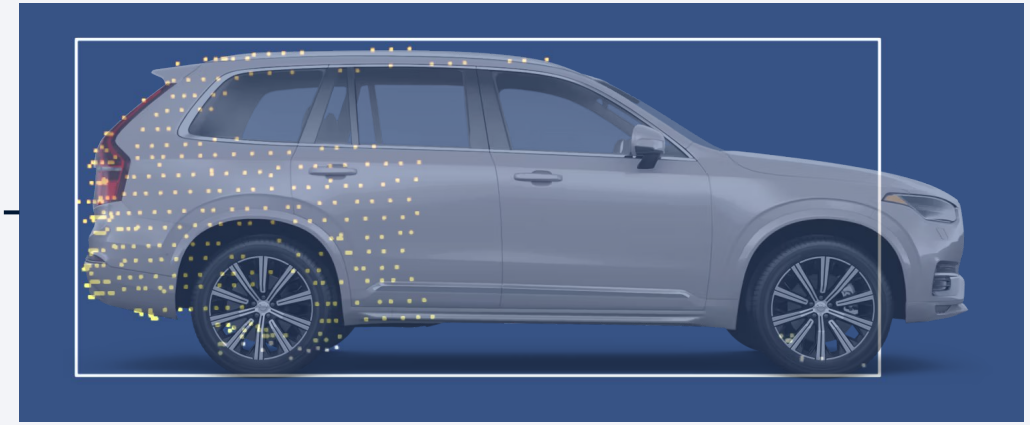
(a) Wider and taller vehicle due to oversized tires and lifted suspension.

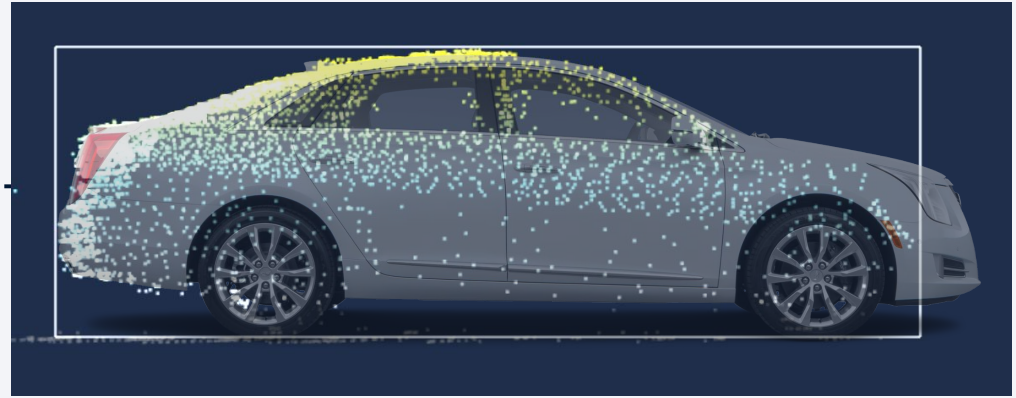
(b) Larger in all three dimensions due to protruding tires and cargo racks.

	%	IoU	Length Error (m)	Width Error (m)	Height Error (m)
All Vehicles	100%	0.883	0.259	0.077	0.075
Modified	17.1%	0.842	0.302	0.119	0.140
Unmodified	82.9%	**0.892**	**0.250**	**0.068**	**0.062**

Surpassing Human Labels

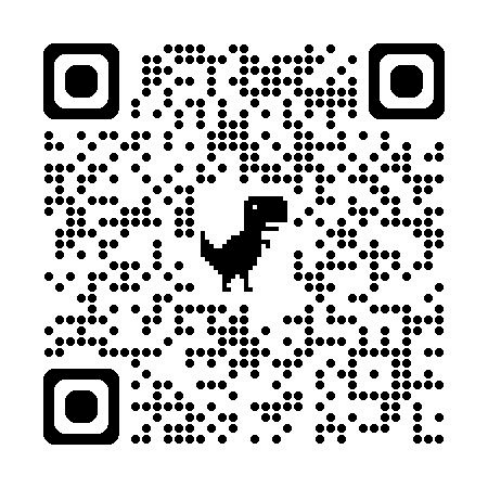






Key Takeaways

- **Reasoning beats guessing:** Explicitly prompting for make/model/generation drastically improves accuracy
- **Vision fills the geometry gap:** 2D semantic understanding can solve 3D lidar occlusion
- **High ROI:** Faster, higher-quality labels with zero-shot inference. No fine-tuning required



https://arxiv.org/abs/2605.21747

Contact: schen@aurora.tech, linkedin.com/in/stevenzchen

Thank you!