

Peer Review

Review of: "OpenAD: Open-World Autonomous Driving Benchmark for 3D Object Detection"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

OpenAD is a benchmark for 3D object detection in open-world driving, and it tries to measure two things at once: domain generalization (a model trained on one dataset, tested on scenes from others) and open-vocabulary recognition (objects whose categories never appeared in training). The authors pick 2,000 scenes from Argoverse 2, KITTI, nuScenes, ONCE, and Waymo, each containing at least one corner-case object the source dataset did not label. They add 6,597 new corner-case objects to 13,164 common objects already annotated in the source datasets, which gives 19,761 objects (§3.1; I checked the sum, $6,597 + 13,164 = 19,761$, and it matches the "19.8K" in Table 1, row "OpenAD (ours)"). Scene discovery and category naming run through a multimodal LLM, with manual checks at each stage (§4). The metrics count a prediction as a true positive only when it passes both a positional threshold and a CLIP-based semantic similarity threshold (§3.2).

On top of the benchmark, the paper offers two baselines. One lifts boxes from any 2D open-world detector into 3D with a small trainable "2D-to-3D Bbox Converter" built on Depth Anything, SAM, and PointNet (§5.1, Figure 4). The other fuses a general and a specialized detector through score alignment and a dual-threshold NMS (§5.2). Tables 2 and 3 evaluate a wide set of 2D and 3D models, and the main message is that open-world models generalize but lack precision, specialized models are precise but blind to unseen classes, and the fusion gets a share of both.

2. Whether the conclusion holds on the data

I think the benchmark itself is a real contribution. Unifying five datasets into one format with a released toolkit removes a lot of dull engineering for anyone doing cross-domain 3D work, and the per-object

seen/unseen flag is a neat design: the same benchmark can be reused whatever training set a model comes from. The evaluation in Tables 2 and 3 is broad, covering 14 rows in the 2D table and 21 in the 3D table, and splitting AR into nuScenes versus other scenes and seen versus unseen categories makes the trade-off between specialization and generality visible in a way single-number leaderboards do not.

I checked the numerical claims in the text against the tables. Most hold. §6.2 says the vision-centric baseline gains "0.58 AP and 6.2 AR" over Find n' Propagate: Table 3, row OpenAD-Y (LC), AP 9.43 against row FnP, AP 8.85, is $9.43 - 8.85 = 0.58$, and $AR\ 25.17 - 18.97 = 6.20$. Correct. What I would add is a sentence on modality. The FnP row is marked "L", the OpenAD-Y row being compared is "LC", so the baseline sees camera images the competitor does not. The camera-only row OpenAD-Y (C) has AP 6.26 and AR 13.89, both below FnP. The fairer reading is that the camera branch plus LiDAR depth refinement beats a LiDAR-only open-vocabulary method; the text should say so.

The ablation claims in §6.3 also match Table 4. Adding the pseudo point cloud to the conv branch moves mAR from 36.8 (row Conv only, MLP) to 46.7 (row Conv + Pseudo Point Cloud, MLP), a gain of 9.9, and replacing the MLP with PCA drops mAR from 45.1 to 27.3. I would like to know what "mAR" is here. It is not defined in §3.2, and the caption says the converter is tested on "2D-3D annotation pairs from OpenAD", which suggests ground-truth 2D boxes as input. If so, Table 4 measures the lifting step in isolation, and its numbers cannot be compared with the AR column of Table 3. That is fine, but it needs saying.

The overfitting argument in §6.2 is the most interesting finding in the paper and also the one resting on the thinnest evidence. The claim is that enlarging the backbone makes BEVStereo and SparseBEV overfit. Table 3 supports it for those two: BEVStereo goes from $AR_{\text{nus}}^{\text{seen}}$ 36.73 to 51.69 while $AR_{\text{others}}^{\text{seen}}$ falls from 16.21 to 13.05, and SparseBEV goes from 60.04 to 61.36 while the out-of-domain column falls from 7.48 to 7.09. But the same table has BEVFormer, ResNet50 to ResNet101-DCN, where the in-domain column rises from 39.38 to 51.86 and the out-of-domain column also rises, from 15.85 to 16.59. So backbone size alone does not predict the drop. The SparseBEV change of 0.39 AR is also small enough that I cannot tell it apart from noise, since every row is a single evaluation with no interval. I would soften the sentence to "in two of three backbone pairs" and add confidence intervals.

That leads to my main statistical point. None of the tables report variance, bootstrap intervals, or repeated runs. Many headline differences are under one AP point: the 0.58 AP above, or Table 3's best ensemble (row OpenAD-Ens, LC, OpenAD-Y + BEVFusion, AP 16.34) against BEVFusion alone (AP 15.57), a gap of 0.77. With 2,000 scenes, a scene-level bootstrap is cheap, and it would show which gaps are real.

The AR gains of the ensemble are large (44.16 versus 33.50 in the same two rows), and I expect them to survive; the AP gains I am less sure about.

Something else I noticed in Table 3: the ensemble's unseen-category recall is copied exactly from its open-world component. Row OpenAD-Ens (LC, OpenAD-Y + BEVFusion) has AR_{nusc}^{unseen} 13.83 and AR_{others}^{unseen} 9.84, identical to row OpenAD-Y (LC). This is what you would expect because BEVFusion contributes 0.00 on unseen classes, and it confirms the fusion NMS does not suppress the open-world boxes. It also means "achieving favorable performance in both seen and unseen" (§6.2) is really "keeping the open-world model's unseen recall while adding the specialist's seen recall". I would phrase it that way because it is an honest and useful result.

The annotation pipeline is where I would ask for the most added detail. §4 reports that the chosen MLLM reaches "approximately 65%" on 30 hand-picked hard samples and "around 90% on the entire data". Thirty samples is a very small set for choosing among GPT-4V, Claude 3 Opus, and InternVL 1.5, and no per-model accuracies are given, only that InternVL won. I could not find how the 90% on the full data was measured, against which reference, or by how many annotators. Since the semantic label is half of the true-positive criterion, the reliability of these labels directly sets the ceiling on every AR number in Tables 2 and 3.

3. Reproducibility

The authors have done the most important thing: data, toolkit, and evaluation code are released at the GitHub link in the abstract. The detector variants used in evaluation are named precisely in §6.1 (OWLv2-CLIP-L/14-ST+FT, YOLO-Worldv2-XL, GLIP-L, GroundingDINO-B, the "vg-grit5m" GenerateU checkpoint), and the backbones of every row are listed in Tables 2 and 3. That is better than most benchmark papers I read.

The gaps are mostly on the MLLM and metric side.

- The CLIP model used for the semantic score is not named. Cosine similarity between text embeddings depends heavily on the checkpoint, and thresholds of 0.5, 0.7, and 0.9 mean different things for different models.
- The MLLM prompts are described only in words ("closed image prompts, such as 2D bounding boxes or circles"). No prompt text, no model snapshot for GPT-4V, no decoding parameters such as

temperature. InternVL is identified as version 1.5, which helps, but the exact checkpoint size is not stated.

- For the 2D-to-3D converter, training details are absent: optimizer, learning rate, epochs, seeds, the PointNet and MLP sizes, and the compute used. §5.1 says only that "only a few parameters" are trainable.
- The fusion in §5.2 says the two models' confidence scores are aligned and that NMS uses IoU and semantic-similarity thresholds, but the alignment rule and both threshold values are not given.
- Figure 2 gives the per-dataset composition of the 2,000 scenes, but as rendered on Qeios, I could not read the counts; the caption says only that ONCE contributes the most scenes.

One question about the data source. §4 says scenes were selected from "the validation and test sets of five autonomous driving datasets," and §3.1 says the 13,164 common objects come from "the annotations of ... the original dataset." For several of these datasets, the test-set labels are not public. I would ask the authors to state which splits each dataset contributed, and where common-object labels for any test-set scenes came from.

4. What specifically to fix

1. §2.1 (3D Benchmark) says OpenAD "contains 206 different corner-case object categories," while §3.1 says all object semantics "can be divided into 206 classes," which would include common objects. Please decide which is true and use it in both places and in Table 1, column Classes.
2. **§6.2, first paragraph.** State that the 0.58 AP / 6.2 AR gain over FnP compares an LC model with an L model (Table 3, column Modality), and report the camera-only comparison too.
3. Define mAR in §3.2 or in the Table 4 caption, and say whether the converter in Table 4 receives ground-truth 2D boxes.
4. Add scene-level bootstrap intervals to the AP and AR columns of Tables 2 and 3. At minimum, for the ensemble rows and their component rows.
5. **§6.2, overfitting paragraph:** mention the BEVFormer ResNet50 to ResNet101-DCN pair (Table 3, $AR_{\text{others}}^{\text{seen}}$ 15.85 to 16.59), which does not follow the stated pattern.
6. In §4, give the per-model accuracy of GPT-4V, Claude 3 Opus, and InternVL 1.5 on the 30-sample set, describe how the "around 90%" figure was measured, and report inter-annotator agreement for the final category labels and for the manually corrected 3D boxes.

7. Name the CLIP checkpoint in §3.2. A small calibration table would help a lot: similarity scores for a few near pairs ("car" / "truck," "cone" / "barrier") and far pairs, so readers can see what 0.5 and 0.9 actually admit. The small nonzero unseen recall of specialized models (for example, Table 2, row MaskRCNN ResNet50, $AR_{\text{others}}^{\text{unseen}}$ 0.07) suggests some unseen tags pass the 0.9 threshold against seen class names; worth a sentence.
8. Publish the MLLM prompts and the decoding settings in an appendix.
9. Explain in §3.1 how free-text natural language tags were grouped into the 206 classes, and whether that grouping was manual.
10. §5.2: give the score-alignment formula and the two NMS thresholds.
11. Specify dataset splits per source in §3.1 (see my question in section 3 above), and include the scene counts from Figure 2 as text or a small table.

5. Recommendation

Minor revisions. This is a useful, carefully assembled benchmark with released code and data, and its central empirical observation, that in-domain gains on nuScenes can hide a loss of generalization, is supported by the tables for most of the model pairs shown. The numbers in the text agree with the tables wherever I checked them. What is missing is mostly description: the reliability of the MLLM-assisted labels, the CLIP model behind the semantic threshold, training details of the converter, and uncertainty estimates for small AP gaps. None of these requires new experiments beyond a bootstrap and an annotation-agreement check, and both would make the benchmark easier to trust and to adopt.

Declarations

Potential competing interests: No potential competing interests to declare.