A. Ablation on the Annotation Pipeline.

As shown in Figure 6, we conduct experiments by employing diverse visual and textual prompts, along with various MLLMs, and select the optimal approach.

The experimental results concerning visual prompts indicate that, as object cues, squares outperform ellipses, while arrows perform less satisfactorily than both. Notably, it is crucial for objects located at the edges of images to maintain the closure of their bounding squares.

Objects like open manholes and wires falling on the road are difficult to identify for existing MLLMs. For such objects, MLLMs tend to respond with other nearby objects. Requiring existing MLLMs to rethink may still not improve the accuracy of their responses.

We use GPT-4V [47], Claude 3 Opus [2], and InternVL 1.5 [13], with InternVL exhibiting the best performance. This may be because InternVL has been trained on more autonomous driving data.

Accuracy is manually calculated based on five repetitions of testing on 30 highly challenging samples. During the manual verification of automated annotations, we conducted a preliminary assessment of the accuracy of the pipeline. The final MLLM and prompt achieve an accuracy rate of approximately 90% on the entire OpenAD data.

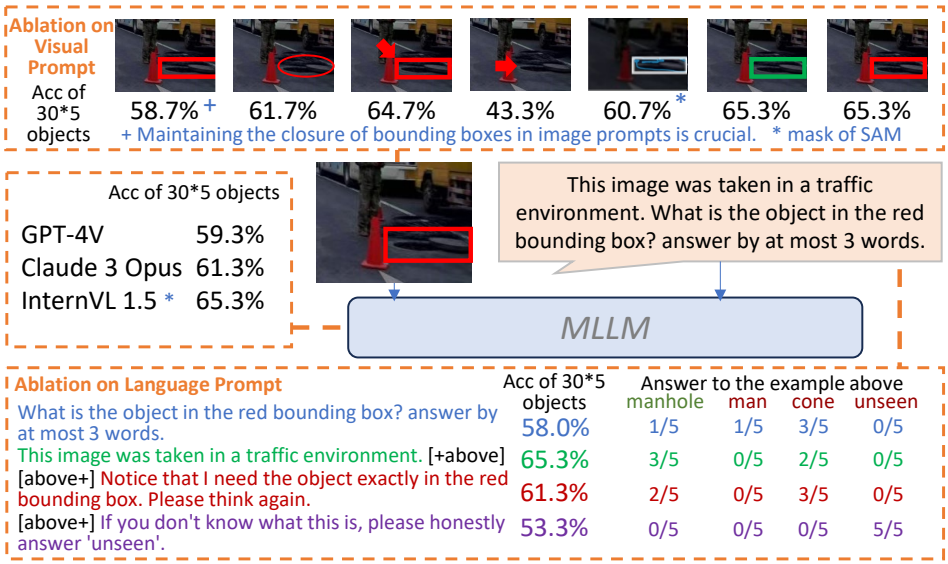

Figure 6: **Ablation on annotation pipeline.** We conduct experiments by employing diverse visual and textual prompts, along with various MLLMs, and select the optimal approach. Accuracy is manually calculated based on five repetitions of testing on 30 highly challenging samples.

B. Further Enhance the Performance of the Proposed Baseline.

B.1 Cross-dataset Training.

Since the proposed converter's training relies on 2D-3D ground truth bounding box pairs, our framework enables convenient cross-dataset training. Table 5 shows that training on three datasets (common classes only) leads to performance gains.

B.2 Using an Instance Segmentation Model.

Some pseudo point clouds generated from background pixels (e.g., road surface within bounding boxes) may introduce noise. To eliminate this noise, we utilize the Segment Anything Model [32]

Table 5: **Performance Comparison: Single Dataset vs. Cross-Dataset Training.** AP and AR of our proposed method can be further improved by training on multiple datasets.

Method	Training on	AP↑	AR↑	ATE↓	ASE↓
OpenAD-G	nuScenes	15.14	34.46	1.056	0.649
OpenAD-G	nuScenes + Waymo + KITTI	19.42	38.08	0.926	0.662
OpenAD-Ens	nuScenes	16.30	48.25	0.858	0.520
OpenAD-Ens	nuScenes + Waymo + KITTI	19.72	53.41	0.869	0.546

Table 6: **Performance Comparison: With vs. Without Segmentation**. This module is trained on nuScenes training set and tested on OpenAD. The 2D proposals are generated by GenerateU [15]. Segmentation results are derived from Segment Anything [32].

Depth	Segmentation	AP	AR
Frozen Depth Anything	✗	9.02	23.32
Frozen Depth Anything	✔	9.07	24.09

(SAM) to segment the object with the 2D box as the prompt, yielding a segmentation mask. While using SAM can bring about marginal improvements, it bloats the framework. Therefore, we have excluded segmentation from the latest version of our baseline. However, if the 2D model used in the framework inherently supports instance segmentation (e.g., VL-SAM [39]), this performance gain can be achieved without additional computational overhead.

C. More Statistics on OpenAD Data.

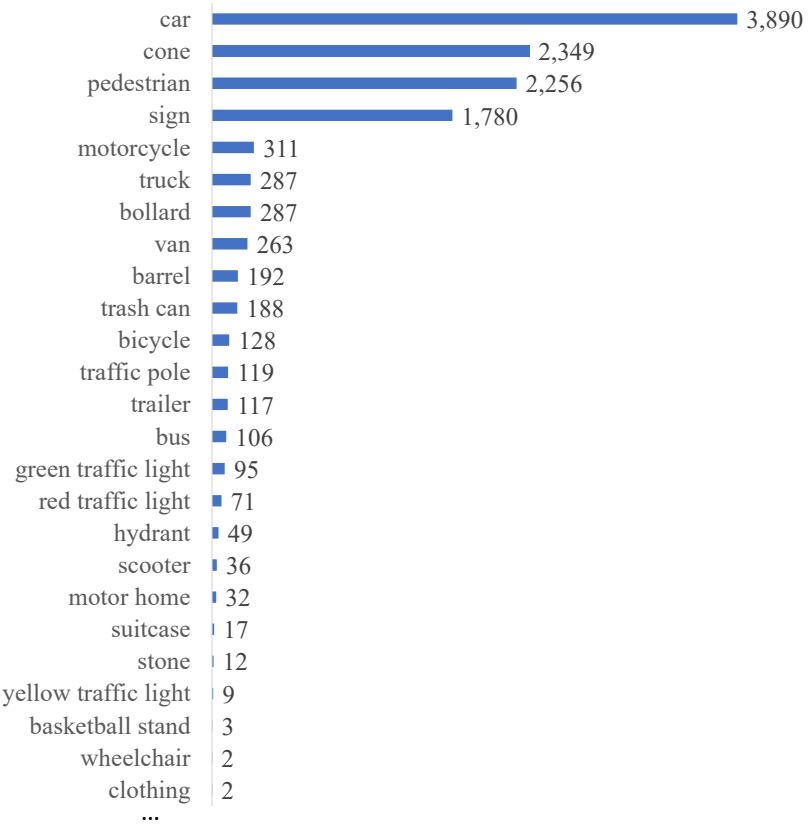

Figure 7: **Statistics on the number of objects in certain categories in OpenAD.**

Since OpenAD is designed to evaluate a model's ability to understand unknown objects, we cannot disclose all category labels in OpenAD. However, we provide quantitative statistics for a subset of labels (common objects or those illustrated in the sample data), as shown in Figure 7. Additionally, Figure 1 demonstrates the diversity of the OpenAD dataset.

D. Broader Impacts Statement.

All data utilized in OpenAD are sourced from published datasets. We do not see potential privacy-related issues. This study may inspire future research on open-world perception models.