FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Towards Open-Vocabulary 3D Object Detection for Autonomous Driving

Pedro Rafael Angélico Madureira



Master in Informatics and Computing Engineering

Supervisor: Prof. Luís Filipe Teixeira

External Supervisor: José Diogo Monteiro

October 1, 2026

© Pedro Rafael Angélico Madureira, 2026

Towards Open-Vocabulary 3D Object Detection for Autonomous Driving

Pedro Rafael Angélico Madureira

Master in Informatics and Computing Engineering

Approved in oral examination by the committee:

President: Prof. André Restivo
Examiner: Prof. António Neves
Supervisor: Prof. Luís Filipe Teixeira

October 1, 2026

Resumo

A Deteção de Objetos (OD) 3D tradicional em Condução Autónoma (AD) baseia-se num conjunto pré-definido de categorias, o que limita a robustez da perceção perante novos obstáculos. Embora o paradigma de Vocabulário Aberto (OV) aborde esta limitação através de Modelos de Visão-Linguagem (VLMs) 2D, a sua adaptação para domínios 3D em ambientes exteriores continua a ser dificultada por vários fatores, incluindo a esparsidade das nuvens de pontos LiDAR, erros de *clustering* heurístico durante a geração *offline* de pseudo-anotações e enviesamentos decorrentes de arquiteturas dependentes de categorias.

Primeiramente, para mitigar os erros de sobressegmentação e subsegmentação inerentes ao *clustering* em LiDAR não restringido, introduzimos uma *pipeline* local de geração de pseudo-anotações. Assim, o nosso método melhora a precisão inicial das pseudo-anotações, ao restringir o *clustering* espacial a *frustums* de câmara 2D derivados do Grounding DINO e ao substituir o *clustering* hierárquico global por uma abordagem Density-Based Spatial Clustering of Applications with Noise (DBSCAN) local. Quando utilizadas na supervisão de detetores, estas anotações resultam em ganhos até $+2.1$ mAP e $+3.8$ NDS relativamente à *baseline* de *clustering* global reproduzida.

Em segundo lugar, propomos o OV-D3D, uma arquitetura que substitui canais multiclasse por um único canal de *objectness* agnóstico à classe. Para compensar a perda dos *priors* dimensionais por categoria, integrámos módulos de escala e de campo recetivo dinâmico. No *benchmark* nuScenes completo, o OV-D3D atinge 28.8 mAP. Isto representa uma recuperação de $+1.0$ mAP face a um controlo simples de um canal, exibindo um compromisso esperado face à *baseline* dependente de categoria (31.1 mAP).

Por fim, sob o nosso Protocolo Holdout Zero-Shot, esta formulação de *objectness* de canal único supera consistentemente a *baseline* na descoberta de novas geometrias no limiar padrão de 2.0 m. Além disso, atinge um recall global no holdout de 76.70% (comparado com os 70.29% da baseline), registando ganhos na descoberta de barreiras, motociclos e atrelados não anotados. Embora a avaliação oficial do *benchmark* resulte em 0 mAP devido ao desalinhamento semântico multimodal, a nossa análise sugere que as representações contínuas preservam a proximidade a sub-caraterísticas visuais. Isto indica que as formulações de canal único localizam formas físicas novas, embora a destilação semântica multimodal permaneça como o principal obstáculo.

Palavras-chave: Vocabulário Aberto • Deteção de Objetos 3D • Pseudo-Anotações 3D • Alinhamento Visão-Linguagem • Condução Autónoma

Abstract

Traditional 3D Object Detection (OD) in Autonomous Driving (AD) relies on closed-set categories, which limits perception robustness when encountering novel obstacles. While the Open-Vocabulary (OV) paradigm addresses this constraint via 2D Vision-Language Models (VLMs), transferring these capabilities to outdoor 3D domains remains hindered by several factors, including LiDAR point cloud sparsity, heuristic clustering errors during offline pseudo-labelling, and category-dependent architectural biases.

First, to overcome the over- and under-segmentation errors inherent to unconstrained LiDAR clustering, we introduce a localised pseudo-label generation pipeline. Thus, our method improves initial pseudo-label precision, by restricting spatial clustering to 2D camera-projected frustums derived from Grounding DINO and replacing global hierarchical clustering with a local Density-Based Spatial Clustering of Applications with Noise (DBSCAN) approach. When used for detector supervision, these cleaner labels yield downstream performance gains of up to $+2.1$ mAP and $+3.8$ NDS over our reproduced global clustering baseline.

Second, we propose OV-D3D, an architecture that replaces multi-class heatmaps with a single class-agnostic objectness channel. To compensate for the loss of categorical dimensional priors, we integrate scale and dynamic receptive-field modules. On the full nuScenes benchmark, OV-D3D achieves 28.8 mAP. This represents a $+1.0$ mAP recovery over a 1-channel control, while exhibiting an expected trade-off against the category-dependent baseline (31.1 mAP).

Finally, under our Zero-Shot Holdout Protocol, this single-channel objectness, at the 2.0 m threshold, consistently outperforms the baseline in discovering novel geometries. Also, it achieves a 76.70% global holdout recall (compared to 70.29% for the baseline), with gains across unannotated barriers, motorcycles, and trailers. Although the official benchmark evaluation yields 0 mAP due to cross-modal semantic misalignment, our analysis suggests that continuous representations preserve proximity to visual sub-features. This indicates that single-channel formulations localise novel physical shapes, leaving cross-modal semantic distillation as the critical bottleneck.

Keywords: Open-Vocabulary • 3D Object Detection • 3D Pseudo-Labelling • Vision-Language Alignment • Autonomous Driving

The United Nations Sustainable Development Goals

The United Nations Sustainable Development Goals (SDG) provide a global blueprint for achieving a sustainable and resilient future. They encompass 17 integrated goals[1] addressing major global challenges, including environmental degradation, human health, technological innovation, and sustainable infrastructure. This section outlines how the research conducted in this dissertation directly contributes toward advancing specific targets within this international framework.

Autonomous Driving (AD) systems have the potential to revolutionise modern transportation by making it safer, more efficient, and highly accessible. However, to operate reliably in dynamic and real-world environments, these systems must be able to accurately perceive and react to a long-tail distribution of rare, unannotated, or unexpected objects. Traditional closed-set 3D object detection models often fail to identify these novel obstacles, potentially misclassifying them into existing base classes or ignoring them as background, which prevents the vehicle from initiating braking and increases the probability of a collision. By proposing a novel architecture, this thesis directly contributes to the safety, innovation, and resilience of next-generation intelligent transport systems. As a result, this dissertation supports the following objectives, especially these particular targets:

SDG 3 Ensure healthy lives and promote well-being for all at all ages.

- **3.6** By 2030[2], halve the number of global deaths and injuries from road traffic accidents.

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation.

- **9.5** Enhance scientific research, upgrade the technological capabilities of industrial sectors in all countries, in particular developing countries, including, by 2030, encouraging innovation and substantially increasing the number of research and development workers per 1 million people and public and private research and development spending.

SDG 11 Make cities and human settlements inclusive, safe, resilient and sustainable.

- **11.2** By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older persons.

[1] United Nations SDGs platform: https://sdgs.un.org/goals, accessed 16 Sep 2026.

[2] Target originally set for 2020, extended to 2030 by the United Nations Decade of Action for Road Safety 2021-2030 (https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/decade-of-action-for-road-safety-2021-2030), accessed 16 Sep 2026.

Table 1: Mapping of Dissertation Contributions to UN Sustainable Development Goals

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.6	By proposing a class-agnostic 3D objectness formulation, this work mitigates the risk of autonomous perception systems ignoring novel physical obstacles as empty space, improving collision avoidance and overall road safety.	• Class-Agnostic Recall across spatial distance thresholds ($\mathbb{D} = \{0.5, 1.0, 2.0, 4.0\}$ m) on unannotated holdout obstacles. • Detection robustness evaluated across cumulative range intervals (0–18 m, 0–34 m, 0–54 m).
9	9.5	Advances open-vocabulary 3D detection by decoupling geometric proposal generation from semantic classification, providing a reproducible and peer-reviewed benchmark methodology for the research community.	• Peer-reviewed research. • Comparative benchmark metrics on the nuScenes dataset (Mean Average Precision (mAP) and nuScenes Detection Score (NDS)) against state-of-the-art baselines.
11	11.2	Improves automated pseudo-labelling and geometric discovery of diverse obstacle classes, facilitating the reliable deployment of autonomous perception within sustainable public transport and shared mobility systems.	• Pseudo-label geometric precision and recall (IoU ≥ 0.5) under camera-LiDAR frustum constraints. • Mean agnostic recall across novel holdout classes.

Acknowledgements

Completing this thesis has been a demanding but incredibly rewarding process, and I certainly did not reach this milestone alone.

First, I thank my supervisors, Luís and José, for their guidance, and, above all, the academic freedom they granted me throughout this project. That trust was instrumental in shaping my research capabilities and independence. I am equally grateful to Filipa for opening the door to the partnership with Bosch and to all my colleagues there, whose collaboration and camaraderie made this experience truly remarkable.

This work was developed under Project ATLAS–Trusted Autonomous Navigation, supported by Portuguese National Funds (COMPETE2030), with Funding Reference COMPETE2030-FEDER-00646700.

Beyond the office, this work stands upon the foundation of the people who have shaped who I am. To my family, my professors, and the friends who walked alongside me on this journey: thank you for challenging me, supporting me, and helping me grow into a better human being.

To Sofia: thank you for sharing this life with me. With you, the high points are always brighter, and the heavy moments become so much easier to bear.

Finally, a very special thank you goes to my father and my grandparents. They taught me the true meaning of resilience by teaching me to never give up, whether faced with a first, a second, or even a thirtieth difficulty. Among them, a very special light belongs to you, Grandma. You were so much more than a grandmother: a teacher, a mentor, and my truest friend. Wherever you are, your love and your lessons will remain a part of all my future accomplishments.

Pedro Angélico

Declaration of honour

I declare that this dissertation is my own work and has not previously been submitted or assessed at this or any other institution. References to other authors (statements, ideas, thoughts), as well as the use of published works, strictly comply with the rules of attribution and are duly identified in the text and in the bibliographic references, in accordance with the applicable referencing standards. I am aware that the practice of plagiarism or self-plagiarism constitutes an academic offence, as established in the U.Porto Code of Ethics.

I further declare that portions of the text in this dissertation are adapted from a peer-reviewed paper titled: *"Decoupling Geometry and Semantics for Robust Open-Vocabulary 3D Object Detection"* [1]. This paper has been accepted for poster presentation at the DriveX workshop of the European Conference on Computer Vision (ECCV) 2026 and is pending formal publication in the conference proceedings. This work, of which I am the first author, was developed in collaboration with both my supervisors. I state with transparency and rigor that all content, implementations, and evaluations included in this dissertation are exclusively my own original work. The express agreement signed by all co-authors validating the independent nature of this dissertation is provided in Appendix F. Because the proceedings are pending publication, formal publisher authorisation is not applicable. However, the conference's standard author rights permit the inclusion of accepted manuscript content in academic theses.

I also declare that I have used Artificial Intelligence tools to aid in text refinement and in code development, which have both been carefully checked for errors, inaccuracies, unfounded attributions or other forms of bias in content and arguments, as well as determining authorship.

Pedro Rafael Angélico Madureira



"Intelligence is what you use when you don't know what to do"

Jean Piaget

Contents

1 Introduction **1**
- 1.1 Context and Motivation . . . 1
- 1.2 Problem . . . 2
- 1.3 Research Questions . . . 2
- 1.4 Contributions . . . 3

2 State of the Art **5**
- 2.1 Foundations of 3D Object Detection in Autonomous Driving . . . 6
  - 2.1.1 LiDAR-based 3D Object Detection . . . 6
  - 2.1.2 Camera-based 3D Object Detection . . . 7
  - 2.1.3 Multimodal-based 3D Object Detection . . . 7
- 2.2 Foundations of Open-Vocabulary Perception . . . 8
  - 2.2.1 Detection through Contrastive Vision-Language Alignment . . . 8
  - 2.2.2 Detection through Self-Supervised Foundations . . . 9
  - 2.2.3 Detection via Foundation Segmentation Models or Generative Models . . . 9
  - 2.2.4 Open-Vocabulary 2D Object Detection . . . 9
- 2.3 Open-Vocabulary 3D Object Detection . . . 10
  - 2.3.1 Geometry-Driven Object Discovery . . . 11
  - 2.3.2 Vision-Driven Object Discovery . . . 13
  - 2.3.3 Orthogonal Approaches . . . 17
- 2.4 Datasets and Evaluation Metrics in Open-Vocabulary Object Detection . . . 18
  - 2.4.1 Autonomous Driving Datasets and the Holdout Paradigm . . . 18
  - 2.4.2 Evaluation Metrics . . . 20
- 2.5 Discussion: The Domain Gap and Baseline Selection . . . 21

3 Methodology **22**
- 3.1 The Baseline Architecture: OV-SCAN . . . 22
  - 3.1.1 Label Generation Pipeline . . . 22
  - 3.1.2 OV-SCAN architecture . . . 25
- 3.2 Baseline Reproduction and Implementation . . . 27
- 3.3 Limitations of the Baseline . . . 30
  - 3.3.1 Limitations of the Pseudo-Label Generator . . . 30
  - 3.3.2 Limitations of the Model . . . 31
- 3.4 Improved Pseudo-Label Generation Pipeline . . . 32
- 3.5 The Proposed OV-D3D Architecture . . . 32
  - 3.5.1 Class-Agnostic Geometric Discovery . . . 34
  - 3.5.2 Vision-Language Alignment . . . 35
  - 3.5.3 Query Decoding and Optimisation . . . 36

3.6 Experimental Setup . . . 36
3.6.1 Datasets . . . 37
3.6.2 Subset Preserving Class Distribution . . . 37
3.6.3 The Zero-Shot Holdout Protocol . . . 37
3.6.4 Training Setup . . . 38
3.6.5 Evaluation Metrics . . . 38

4 Results **40**
4.1 Establishing the Baseline . . . 40
4.1.1 Validating the Codebase and Evaluation Metrics . . . 40
4.1.2 Validating the Training and Label Generation Pipeline . . . 41
4.2 Evaluation of the Improved Pseudo-Label Generation . . . 43
4.2.1 Pseudo-Label Geometric Quality . . . 43
4.2.2 Qualitative Clustering Analysis . . . 44
4.2.3 Computational Viability . . . 44
4.2.4 Downstream 3D Detection Impact . . . 45
4.3 Our Improved Architecture: OV-D3D . . . 46
4.3.1 Macro-Level 3D Detection Performance . . . 46
4.3.2 Robustness to Spatial Degradation at Range . . . 48
4.3.3 The Zero-Shot Paradox . . . 48
4.3.4 Isolating Geometric Localisation . . . 49
4.3.5 The Semantic Alignment Bottleneck . . . 50
4.4 Ablation Studies . . . 53
4.4.1 Impact of Prompt Engineering Strategies . . . 53
4.4.2 Sensitivity of the 3D Pipeline . . . 55
4.4.3 OV-D3D Component Ablations . . . 57

5 Conclusion **61**
5.1 Answers to the Research Questions . . . 61
5.2 Limitations and Future Work . . . 64

References **66**

A Systematic Literature Review **72**
A.1 Search Queries and Initial Results . . . 72
A.2 Screening and Exclusion Criteria . . . 73
A.2.1 Automatic Filtering (Pre-Screening) . . . 73
A.2.2 Manual Screening and Eligibility . . . 75

B Baseline Reproduction **76**
B.1 Structure of the OV-SCAN Reference Labels . . . 76
B.2 Subset Selected to Preserve Class Distribution . . . 76
B.3 Extended Grid Search Results for Proposal Generation Reproduction . . . 78

C Implementation Hyperparameters **80**
C.1 Pseudo-Label Generation Pipeline . . . 80
C.2 OV-D3D Training Configuration . . . 81

D Extended Metrics and Semantic Analysis **82**
D.1 Full Class-Agnostic Recall Results 82
D.2 Full Semantic Confusion Matrices 82

E Preliminary Explorations and Paradigm Limits **86**
E.1 Temporal Tracklet Optimisation 86
E.2 Anchor-Free Optimisation on Sparse Point Clouds 87

F Paper Inclusion Authorisation **89**

List of Figures

2.1 Publication Trend by Research Query 6
2.2 Representative Geometry-Driven 3D Open-Vocabulary Object Detection (OV-OD) Architecture 12
2.3 Representative Vision-Driven 3D OV-OD Architecture via Frustum Lifting . . . 14
2.4 Representative Vision-Driven 3D OV-OD Architecture via Explicit Box Learning 16
2.5 Statistical overview of the OpenAD benchmark 19

3.1 Overview of the OV-SCAN architecture 23
3.2 Step-by-Step Visualisation of the Proposed Pseudo-Label Generation Pipeline . . 33
3.3 Overall Architecture of OV-D3D 34

4.1 Validation of 2D Visual Proposals 42
4.2 Class-Wise Label Generation Volumes 43
4.3 Comparative BEV Analysis of Vehicle Bounding Box Optimisation 45
4.4 Zero-Shot Analysis of a Novel `Motorcycle` Instance 49
4.5 Raw Cosine Similarity Semantic Confusion Matrices (Full Benchmark Protocol) 51
4.6 Qualitative Impact of 2D Prompt Engineering Strategies 54
4.7 Qualitative BEV Comparison of 3D Region Proposals 56

A.1 Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 Flow Diagram 74

D.1 Softmax Probability Semantic Confusion Matrices (Full Benchmark Protocol) . . 84
D.2 Raw Cosine Similarity Semantic Confusion Matrices (Holdout Dataset) 84
D.3 Softmax Probability Semantic Confusion Matrices (Holdout Dataset) 85

E.1 Comparative BEV Trajectory Analysis Over Sequential Frames 87

F.1 Authorisation for Inclusion of Co-Authored Paper Content 89

List of Tables

1 Mapping of Dissertation Contributions to UN Sustainable Development Goals . . iv

2.1 Comparison of Standard AD Datasets and Benchmarks 19

4.1 Validation of the Reverse-Engineered Evaluation Script 41
4.2 Validation of Training and Label Generation Pipelines 42
4.3 Overall Pseudo-Label Localisation Quality 43
4.4 3D Detection Performance: Reproduced Global Labels vs. Local Labels 45
4.5 Class-Wise mAP Shifts: Global vs. Local Labels 46
4.6 3D Object Detection Performance on nuScenes Validation Set 47
4.7 Robustness Evaluation Across Distance Intervals 48
4.8 Class-Agnostic Recall on Zero-Shot Holdout Dataset 50
4.9 Quantitative Comparison of 2D Prompt Engineering Strategies 54
4.10 Sensitivity Summary of 3D Pipeline Hyperparameters 55
4.11 Impact of Particle Swarm Optimisation (PSO) Iteration Count on Bounding Box Quality 56
4.12 Impact of Temporal Tracking Configurations on Detection mAP 57
4.13 Architectural Component Ablations of OV-D3D 58
4.14 Ablation Evaluation Across Distance Intervals 59

A.1 Initial Literature Search Results 73

B.1 Baseline Performance on Stratified Training Subsets 78
B.2 Comprehensive 3D Pipeline Grid Search Results 79

C.1 Hyperparameters for 2D Proposal Generation (Grounding DINO) and Cross-Modal Association. 80
C.2 Point Cloud Clustering configurations 80
C.3 (PSO) hyperparameters 81
C.4 Loss weighting coefficients and network architecture parameters for the OV-D3D model. 81

D.1 Full Class-Agnostic Recall Across All Categories 82
D.2 Class-Agnostic Recall Across Distance Thresholds 83

List of Acronyms

AD	Autonomous Driving
A-OSE	Absolute Open-Set Error
AP	Average Precision
AR	Average Recall
ASPP	Atrous Spatial Pyramid Pooling
BEV	Bird's Eye View
CMDB	Cross-Modal Distillation Block
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FFN	Feed Forward Network
H2SA	Hierarchical Two-Stage Alignment
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
IoU	Intersection-over-Union
MAE	Mean Absolute Error
mAOE	Mean Average Orientation Error
mAP	Mean Average Precision
mAR	Mean Average Recall
mASE	Mean Average Scale Error
mATE	Mean Average Translation Error
MLLM	Multimodal Large Language Model
MLP	Multi-Layer Perceptron
MoE	Mixture-of-Experts
NDS	nuScenes Detection Score
NMS	Non-Maximum Suppression
OD	Object Detection
OV	Open-Vocabulary
OV-OD	Open-Vocabulary Object Detection
PCA	Principal Component Analysis
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PSO	Particle Swarm Optimisation
ROI	Region of Interest

RPN	Region Proposal Network
SA	Self-Attention
SAM	Segment Anything Model
SC-NOD	Semantically-Consistent Novel-Object Discovery
SDG	Sustainable Development Goals
VLM	Vision-Language Model
VRU	Vulnerable Road User

Chapter 1

Introduction

1.1 Context and Motivation

Autonomous Driving (AD) systems rely fundamentally on perception to understand their surroundings, including the existence, type, and location of relevant objects in complex and dynamic environments. Therefore, the reliability of this layer is critical, as its outputs directly inform downstream modules responsible for prediction, planning, and control [2]. Traditionally, these perception pipelines are built upon Object Detection (OD) models, a field that has evolved from 2D image-based detection to 3D spatial reasoning to provide the depth and orientation necessary for safe navigation.

Despite significant advances, most state-of-the-art 3D detectors operate under a closed-set assumption, being trained and evaluated on a fixed set of predefined categories [13]. This paradigm restricts their adaptability in real-world scenarios, where previously unseen or rare objects that fall outside the training distribution may appear [5, 65]. If the system misclassifies a novel obstacle as a base class or ignores it as "background," the downstream planning layer will receive inaccurate semantic information, posing a significant safety risk in the context of AD [21]. To reduce this dangerous dependence on exhaustive, predefined label sets, the field is increasingly shifting towards the Open-Vocabulary Object Detection (OV-OD) paradigm [5].

Modern formulations of the Open-Vocabulary (OV) paradigm address these limitations by leveraging vision-language alignment to recognise categories defined by natural language descriptions [53, 76]. While the integration of Vision-Language Models (VLMs), such as CLIP, has accelerated progress in 2D OV-OD, extending this reasoning to 3D is more complex due to the scarcity of annotated 3D data. Currently, much of this research is confined to indoor environments, where RGB-D sensors provide dense geometry and stable pixel-to-point alignment [42, 59]. However, these controlled settings do not reflect the scale, sensor noise, or long-range uncertainty inherent to the world of AD, where outdoor LiDAR is radially degrading and highly sparse at range.

While road safety motivates this research, the experimental scope of this dissertation focuses specifically on the 3D perception layer, evaluating geometric object discovery and vision-language alignment within an offline benchmark.

1.2 Problem

Although recent research has shown promising progress in 2D OV recognition, transferring these capabilities to 3D AD remains an underexplored challenge. In contrast to indoor settings, outdoor perception typically relies on heterogeneous sensing configurations, which may include cameras, LiDAR, or their combination, each exhibiting distinct failure modes and uncertainty characteristics.

While cameras provide rich semantic information suitable for vision-language alignment, LiDAR point clouds are sparse and non-uniformly distributed. This asymmetry introduces alignment challenges, such as projecting a single 3D point onto a semantically diverse image region [34]. These challenges are further intensified by environmental factors such as motion blur, occlusions, and atmospheric interference (e.g., rain or fog), which degrade the reliability of these modalities [63].

Consequently, the architectural assumptions of indoor 3D OV-OD, such as dense geometric cues or stable cross-modal correspondences, fail in the driving context. This mismatch can lead to cascading errors during the generation of offline pseudo-labels. When generating these labels, attempting to group sparse and range-degraded LiDAR points using traditional 3D spatial clustering heuristics often results in over- or under-segmentation, producing corrupted geometric priors for the network to learn from.

Furthermore, current 3D OV-OD architectures targeted for AD [11, 15] still rely on closed-set mechanisms to discover objects. By using predefined categorical heatmaps to locate targets, these models become inherently biased towards known shapes. This restricts their capacity to localise novel objects [1].

Finally, the evaluation of these architectures presents a critical gap. Although several works have recently proposed methods for 3D OV-OD [7, 11, 15, 43], current benchmarks frequently report performance on novel objects to which the model was actually exposed during training through pseudo-labels. Consequently, it remains unclear how well these architectures generalise to geometries they have not been supervised on in real-world driving scenarios.

1.3 Research Questions

To systematically investigate the core bottlenecks of outdoor 3D OV-OD mentioned in Section 1.2 about unreliable pseudo-label generation, architectural closed-set bias, and benchmark evaluation leakage, this thesis formulates three targeted research questions.

As established in Section 1.2, the absence of native 3D foundation models forces current pipelines to depend on unconstrained 3D spatial clustering, which frequently corrupts training targets via over- and under-segmentation. Because 2D foundation models provide mature spatial and semantic representations, a direct opportunity arises to use camera information to guide these 3D heuristics.

RQ 1: How can information from 2D foundation models be used to constrain the heuristic errors inherent to LiDAR point cloud clustering, thereby improving the generation of reliable

3D pseudo-labels?

However, high-quality pseudo-labels address only the data dimension. As highlighted in Section 1.2, leading detectors rely on intermediate closed-set classification to discover candidate objects. Conditioning localisation on predefined classes risks over-specialising the network to known objects, directly conflicting with OV discovery. This motivates exploring alternative formulations that do not rely on those priors. Therefore, it is important to determine whether removing these intermediate supervision signals compromises the detector's baseline fidelity on familiar objects.

RQ 2: What is the impact on the detection performance of base classes when eliminating the network's reliance on intermediate closed-set classification?

Finally, the ultimate promise of the OV paradigm is the reliable detection of objects from both seen and previously unseen classes. Yet, as identified in Section 1.2, prevailing benchmarks fail to provide a true zero-shot evaluation because training splits frequently leak novel categories through pseudo-labels. Consequently, it is pertinent to evaluate 3D OV-OD architectures under an explicit holdout protocol, where class supervision is entirely withheld during training although sensor returns from novel objects remain in the background.

RQ 3: To what extent can 3D OV-OD architectures generalise to novel object geometries in real-world driving scenarios?

1.4 Contributions

To investigate the research questions defined in Section 1.3, the primary contributions of this thesis are summarised as follows:

1. **Experimental Reproduction and Baseline Infrastructure:** At the time this thesis was written, the state-of-the-art OV-SCAN detector only provided its model architecture source code alongside a pre-computed database of cross-modal pseudo-labels, leaving the offline label generation pipeline code closed-source. To establish a controlled baseline for data-generation improvements, we reverse-engineered this offline pipeline from the published literature to create our Global HDBSCAN baseline control. When supervised using the authors' reference labels, our reproduced training pipeline achieved 30.9 mAP on the full nuScenes benchmark (within 0.2 mAP of the published 31.1 mAP), validating the detector training codebase independently of data-preparation variations. Training the model with our reproduced global clustering pipeline yielded 26.7 mAP, establishing an empirical baseline to isolate the impact of subsequent pseudo-labelling modifications.

2. **Improved Pseudo-Label Generation Pipeline:** Whereas the baseline offline pipeline applies HDBSCAN globally across the full LiDAR point cloud, we introduced a localised pipeline that restricts spatial clustering to 2D Grounding DINO camera frustums coupled with a local

DBSCAN implementation, which increased pseudo-label precision against nuScenes ground truth (3D IoU ≥ 0.5) from 28.0% to 34.9% (+6.9%) and recall from 11.1% to 13.1% (+2.0%), translating into downstream gains of +2.1 mAP (to 28.8 mAP) for OV-SCAN and +1.0 mAP (to 27.4 mAP) for OV-D3D under Proposed Local Labels on the full nuScenes benchmark.

3. **Architectural Proposal (OV-D3D):** Whereas prevailing query-based 3D detectors rely on multi-channel categorical heatmaps to initialise proposals, we proposed OV-D3D by collapsing proposal discovery into a single class-agnostic objectness channel complemented by Gated Atrous Spatial Pyramid Pooling (ASPP) and Harmonic Scale Embeddings, achieving 27.8 mAP with the minimal 1-channel control and 28.8 mAP with the complete OV-D3D architecture on the full nuScenes benchmark under Reference Pseudo-Labels (compared to 31.1 mAP published and 30.9 mAP reproduced for the multi-channel baseline).

4. **Empirical Knowledge on Zero-Shot Generalisation and Alignment Bottlenecks:** Open-vocabulary driving benchmarks frequently expose models to novel categories via pseudo-labels during offline training, obscuring true zero-shot capability. We formulated a Zero-Shot Holdout Protocol that excludes target novel classes from supervisory pseudo-labels. The evaluation revealed an empirical divergence: while under the official nuScenes protocol classification yielded 0 mAP for all evaluated architectures, the single-channel objectness formulation demonstrated superior class-agnostic proposal discovery on unannotated geometries. Evaluated at the standard 2.0 m centre-distance matching threshold, OV-D3D substantially outperformed the multi-channel baseline across all holdout classes, securing 65.61% vs. 41.71% on unannotated barriers, 87.61% vs. 77.66% on motorcycles, and 30.77% vs. 23.01% on trailers (with a global dataset recall of 76.70% vs. 70.29%). Diagnostic analysis of unscaled cosine similarities confirmed that continuous latent representations maintain proximity to visual sub-features, suggesting that cross-modal semantic distillation and contextual co-occurrence, rather than geometric proposal discovery, represent the primary bottleneck in zero-shot generalisation.

Chapter 2

State of the Art

The transition of autonomous driving perception from rigid, predefined categories to Open-Vocabulary (OV) understanding requires a fundamental shift in how 3D Object Detection (OD) is formulated. To systematically contextualise the contributions of this thesis, this chapter presents a critical review of the literature spanning three core domains. First, we examine the mature foundations of traditional closed-set 3D OD to establish the architectural baselines (Section 2.1). Second, we review the emergence of the OV paradigm, focusing on how 2D Vision-Language Models (VLMs) enable zero-shot semantic reasoning (Section 2.2). Finally, we discuss the 3D Open-Vocabulary Object Detection (OV-OD) methods, categorising current approaches in different families and identifying the critical domain gaps that motivate this work.

To structure the literature review, the studies examined in this chapter were selected following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines. The search strategy targeted four major academic databases: IEEE Xplore, ACM Digital Library, Scopus, and arXiv. The goal was to capture both peer-reviewed publications and the latest state-of-the-art preprints. The queries were designed to intersect three core domains: traditional closed-set 3D OD to establish architectural baselines, the general OV paradigm to understand VLMs like CLIP [53], and their specific intersection in Autonomous Driving (AD). After applying exclusion criteria regarding temporal relevance, citation impact, and domain applicability, a final set of 21 core studies was selected for deep qualitative synthesis. The complete PRISMA flowchart, exact search queries, and explicit exclusion criteria are detailed in Appendix A.

Crucially, a bibliometric analysis of the initial search results provides quantitative justification for the focus of this thesis. As illustrated in Figure 2.1, while traditional closed-set 3D OD (Q1) exhibits a highly mature publication volume, and general OV methods (Q2) demonstrate explosive growth following the release of foundational models in 2021, the trend line for their specific intersection in 3D AD environments (Q3) remains underexplored. This scarcity of literature quantitatively confirms the novelty of applying OV methods to 3D outdoor perception.

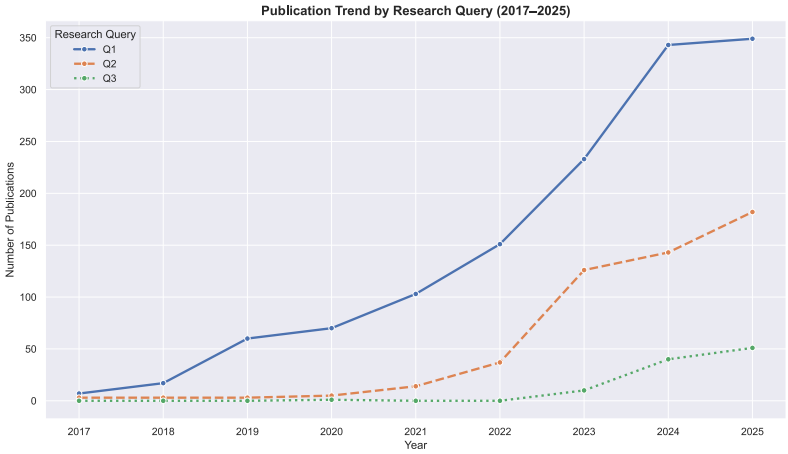


Figure 2.1: Publication Trend by Research Query (2017–2025). The divergence in volume highlights the gap between established paradigms and the emerging intersection of 3D OD and OV methods.

2.1 Foundations of 3D Object Detection in Autonomous Driving

3D Object Detection (OD) aims to localise and classify objects in three-dimensional space by predicting orientated 3D bounding boxes. A 3D bounding box can be parametrised as

$$B = [x_c, y_c, z_c, l, w, h, \theta, class],$$

where (x_c, y_c, z_c) denotes the centre of the cuboid, (l, w, h) its length, width, and height, θ represents the yaw angle, and *class* corresponds to the semantic category of the object. Under the assumption that objects lie on the ground plane, only the yaw angle is required to describe the orientation of the bounding box [45].

Additionally, in a traditional closed-set supervised setting, we assume a dataset $\mathscr{D}$ associated with a fixed and predefined set of object classes $\mathscr{C}$. The dataset is split into training and testing subsets, $\mathscr{D}_{\text{train}}$ and $\mathscr{D}_{\text{test}}$, respectively, with both splits sharing the same class vocabulary $\mathscr{C}$. Consequently, models are trained and evaluated under the assumption that every test instance belongs to one of the classes observed during training.

As discussed before, closed-set 3D OD is a mature research field, widely explored in AD perception systems where accurate understanding of the environment is critical for downstream planning [45]. In this section, we focus on three major families of methods: LiDAR-based, camera-based, and multimodal-based approaches.

2.1.1 LiDAR-based 3D Object Detection

Point-based Methods. Point-based 3D object detection methods aim to detect objects directly from raw point clouds [52, 57]. In these approaches, the input point cloud is processed by a point-based backbone network, where points are progressively sampled and local features are learnt using

point cloud operators. Based on the resulting downsampled points and their associated features, 3D bounding boxes are then predicted. While these methods preserve fine-grained geometric information, point sampling often becomes a computational bottleneck during inference [45].

Grid-based Methods. These methods transform point clouds into grid representations, such as voxels, pillars, or Bird's Eye View (BEV) feature maps. Once converted into grid form, features are extracted using either 2D convolutional neural networks or 3D sparse convolutional networks, depending on the representation. Three main types of grid representations are commonly used:

- **Voxels:** The space is discretised into a regular three-dimensional grid, where each grid cell corresponds to a voxel that is considered non-empty if it contains at least one point [71, 74, 84]. Since point clouds are inherently sparse, most voxels remain empty. In practice, only non-empty voxels are stored and processed for feature extraction [45, 71].
- **Pillars:** The point cloud is discretised only along the horizontal plane, forming a 2D grid in the (x, y) dimensions [29]. All points falling within the same cell are grouped into a vertical column. This representation enables efficient use of 2D convolutional networks while preserving vertical information through feature encoding.
- **BEV feature maps:** Dense 2D representations in which each pixel corresponds to a specific region in the ground plane and encodes aggregated point features within that region. They can be obtained by projecting voxel- or pillar-based features onto the BEV, or directly from raw point clouds [45].

A central trade-off across all grid-based representations lies in grid resolution: finer cells retain geometric fidelity at the expense of memory footprint and inference latency [45].

2.1.2 Camera-based 3D Object Detection

These methods rely solely on visual information to perform 3D OD. Monocular 3D OD does it by relying on a single image [27, 47, 81]. They are particularly challenging due to the lack of explicit depth information, especially when only 2D bounding box annotations are available during training [27, 45, 47]. Nevertheless, monocular approaches can benefit from advances in 2D OD architectures.

More recently, multi-view 3D OD leverages images captured from multiple viewpoints to improve depth estimation and spatial reasoning [23, 35, 67]. Therefore, these methods can alleviate some of the depth ambiguity present in monocular approaches and achieve more accurate 3D localisation.

2.1.3 Multimodal-based 3D Object Detection

These methods combine information from multiple sensor modalities, most commonly cameras and LiDAR [9, 41, 51]. Given the complementary nature of both sensors, where cameras provide rich

semantic information and LiDAR sensors offer accurate depth measurements and precise 3D spatial structure, these methods achieve competitive results when compared with isolated sensors [9, 41, 45, 51].

However, the fusion process introduces additional computational overhead and increases inference latency [9, 41, 45].

2.2 Foundations of Open-Vocabulary Perception

The OV paradigm relaxes the assumption held by the closed-set setting, where both training and testing subsets, $\mathscr{D}_{\text{train}}$ and $\mathscr{D}_{\text{test}}$, respectively, share the same class vocabulary $\mathscr{C}$, by partitioning the class vocabulary into two disjoint subsets: a set of base classes $\mathscr{C}_b$, which are available during training, and a set of novel classes $\mathscr{C}_n$, which are not observed during training. Formally, the training set contains annotations only for objects $x \in \mathscr{C}_b$, while evaluation is performed over the expanded test vocabulary:

$$\mathscr{C}_{\text{test}} = \mathscr{C}_b \cup \mathscr{C}_n$$

In the context of autonomous driving perception, this setting is particularly challenging due to the long-tail distribution of object categories encountered in real-world traffic, where rare, unusual, or safety-critical objects may appear infrequently or not at all in the training data. As a result, evaluation protocols must go beyond exact label matching and instead account for semantic similarity and generalisation to unseen categories, as further discussed in Section 2.4.

This section reviews the 2D foundations that enable semantic and geometric reasoning in 3D OV-OD.

2.2.1 Detection through Contrastive Vision-Language Alignment

VLMs are typically trained using the vast amount of large-scale image pairs available on the internet with the objective of aligning both visual and text embeddings within a common embedding space [25, 33, 53, 82]. CLIP follows this paradigm through a contrastive learning objective during pretraining on 400 million image-text pairs. Given a batch of paired images and textual descriptions, the model jointly optimises an image encoder and a text encoder to maximise the similarity between matching pairs and minimise it for non-matching ones. This strategy enables CLIP to learn semantic associations without relying on task-specific annotations or a fixed set of object categories.

As a consequence, these models can be used for zero-shot classification. At inference time, the text encoder is used to generate embeddings of each textual description or label from a given dataset. Additionally, an image is passed to the image encoder, and then the similarity between each pair is computed, and the most similar pair is selected. Following a similar architecture, ALIGN [25] confirms the generality of contrastive vision-language alignment by extending the paradigm to larger and noisier web-scale datasets.

This design pattern directly inspires geometry-driven 3D OV-OD approaches [7, 10, 42, 61, 78]: once a class-agnostic 3D proposal is generated from a point cloud, it can be projected into the image plane and semantically classified using a frozen VLM.

2.2.2 Detection through Self-Supervised Foundations

Self-supervised visual representation learning aims to extract rich visual features without human-provided labels or language supervision. DINO [8] introduced a self-distillation framework that offers clear separation of object boundaries as one of its main advantages. DINOv2 [50] scaled this approach to a curated dataset of 142 million images and added a masked image modelling objective. It achieves a stronger transfer performance than CLIP-based features on dense prediction tasks, such as semantic segmentation and depth estimation [50]. This highlights a key trade-off between model families: contrastive VLMs provide strong semantic alignment but relatively coarse spatial features, whereas self-supervised models yield spatially precise and object-aware representations without explicit semantic grounding [8, 50, 54].

In 3D OV-OD pipelines, these models are therefore frequently used as visual backbones for feature extraction [15, 42, 73]. However, their lack of intrinsic language alignment requires complementary semantic supervision from VLM.

2.2.3 Detection via Foundation Segmentation Models or Generative Models

Some foundation models do not perform detection directly but provide complementary capabilities that improve object localisation or enable modality bridging. Segment Anything Model (SAM) [26] is capable of generating masks from points, or boxes, without requiring category supervision. Although SAM does not encode semantic information by itself, it is frequently combined with VLMs to associate segmentation masks with textual concepts, as explored in Semantic-SAM [32] and Grounded-SAM [55]. In 3D OV-OD approaches, such models are used to generate masks that reduce background interference and improve the quality of pixel-to-point correspondences prior to 3D lifting or projection [22, 24].

Generative models have also been explored as a means to bridge missing modalities. ControlNet [80], a conditional diffusion framework, enables fine-grained control over image generation based on structural inputs, such as depth maps. In the context of 3D OV-OD, ControlNet [80] is not used for detection itself but rather to hallucinate image representations from sparse 3D observations, as demonstrated in ImOV3D [72]. While promising, such approaches introduce additional uncertainty and computational overhead.

2.2.4 Open-Vocabulary 2D Object Detection

2D OV-OD aims to extend conventional detectors beyond a fixed category set by leveraging language supervision or pretrained vision–language representations. Existing approaches can be broadly grouped according to how semantic knowledge from large-scale vision–language models is integrated into the detection pipeline.

Early methods, such as ViLD [18] and Detic [83], follow a two-stage detection paradigm and rely on knowledge distillation from a pretrained vision–language classifier into a region-based detector. While effective, these approaches require careful training procedures and additional supervision to transfer semantic knowledge reliably.

More recent works simplify this process by freezing a pretrained VLM backbone and training only lightweight detection heads. F-VLM [28] and RegionCLIP [82] demonstrate that frozen CLIP features retain sufficient locality for region-level classification, enabling efficient open-vocabulary detection without explicit distillation. These methods significantly reduce training complexity, but their spatial precision remains limited by the granularity of VLM features.

An alternative formulation recasts detection as language grounding. GLIP [33] and Grounding DINO [39] unify detection and phrase grounding by conditioning bounding box prediction directly on textual queries. This paradigm yields strong open-set performance and has become a popular proposal generator for downstream 3D OV-OD pipelines. However, grounding-based models incur higher computational costs and are sensitive to image degradation, which can be problematic in safety-critical driving scenarios.

Beyond generic benchmarks, a small number of methods explicitly target autonomous driving. z-Prod [58] tackles OV perception in AD by recasting detection as zero-shot discovery of out-of-distribution objects. Using prompt-based vision–language reasoning, it filters out known road participants to reveal rare or safety-critical obstacles without requiring domain-specific supervision. Cas-OVD [16] adopts a camera-only formulation and introduces a multi-refined Region Proposal Network (RPN) to improve recall for small and rare objects, which preserves all anchors during proposal generation and re-weights them instead of hard-filtering. In addition, the method explicitly emphasises moderately difficult samples during training and enriches textual descriptions beyond raw class names, improving semantic discrimination for rare objects.

2.3 Open-Vocabulary 3D Object Detection

These approaches emerge as a response to the fundamental mismatch between closed-set 3D detectors and the open-world nature of autonomous driving scenarios. A 3D OV-OD method aims to localise objects in three-dimensional space while assigning semantic labels that are not restricted to a predefined set of categories, including classes not explicitly observed during training [76]. Unlike traditional 2D OD, 3D OD produces orientated bounding boxes that encode spatial extent, depth, and orientation, which are critical for downstream tasks in AD such as motion prediction and planning. These boxes may be represented directly in 3D point clouds, typically acquired from LiDAR sensors, or projected onto 2D image planes using camera calibration parameters.

A fundamental challenge in this domain is the scarcity of large-scale annotated 3D datasets, as acquiring and labelling 3D data is significantly more costly than in the 2D case [2]. While 2D vision benefits from the availability of massive image-text corpora that enable strong VLMs, such as CLIP, the limited availability of paired 3D-text data makes direct training of 3D-language models

currently impractical. As a result, most existing approaches leverage 2D visual semantics as an intermediate supervisory signal to enable OV reasoning in 3D.

Based on their assumptions regarding supervision and the dominant modality used for object discovery, existing 3D OV-OD methods can be broadly grouped into three architectural families:

- **Geometry-Driven Object Discovery:** Rely primarily on LiDAR data and class-agnostic 3D proposal generation, learning objectness from geometric cues before aligning proposals with 2D vision-language embeddings for semantic assignment.
- **Vision-Driven Object Discovery:** Assume mature 2D OV-OD models as the main source of object discovery, using image-based detections to guide or condition 3D localisation through lifting or geometry-aware regression.
- **Orthogonal 3D OV Approaches:** Encompass methods that do not conform neatly to either paradigm.

2.3.1 Geometry-Driven Object Discovery

Geometry-driven methods assume that geometric cues coming from LiDAR point clouds are enough to learn objectness without category supervision. As illustrated in Figure 2.2, these architectures use a class-agnostic 3D proposal generator, typically a Point Cloud-based RPN, to discover potential objects in 3D space.

This is done by processing the LiDAR point cloud to output candidate 3D bounding boxes. To achieve OV recognition, in many approaches, these geometric proposals are projected onto the image plane via camera extrinsic and intrinsic matrices. The corresponding image crops are processed by a frozen VLM image encoder, while textual queries are processed by the VLM text encoder. Finally, semantic assignment is performed via similarity matching within a shared latent space.

OV-3DETIC [42] uses a teacher-student distillation framework. A 2D detector, enriched by ImageNet-1K [56], generates pseudo-labels for projected 3D proposals, supervising the model on categories beyond the standard 3D training set. While this addresses long-tail distribution issues, empirical results also highlight a performance disparity based on object scale, where the system demonstrates high accuracy for distinct structures like toilets and bathtubs but continues to struggle with small-scale or low-point-density objects such as microwaves and lamps.

CoDA [7] aims to mitigate the limitations of discrete pseudo-labelling by introducing a continuous discovery-driven alignment mechanism. By directly associating 3D geometric features with high-dimensional VLM embeddings, CoDA enables detection via arbitrary textual prompts. This design improves generalisation to novel categories. However, CoDA remains vulnerable to modality misalignment.

UP-VL [48] expands the geometry-driven paradigm by leveraging temporal cues rather than relying strictly on static spatial clustering. It introduces an unsupervised multi-modal auto-labelling pipeline that computes Neural Scene Flow across sequential LiDAR frames, using the motion of 3D

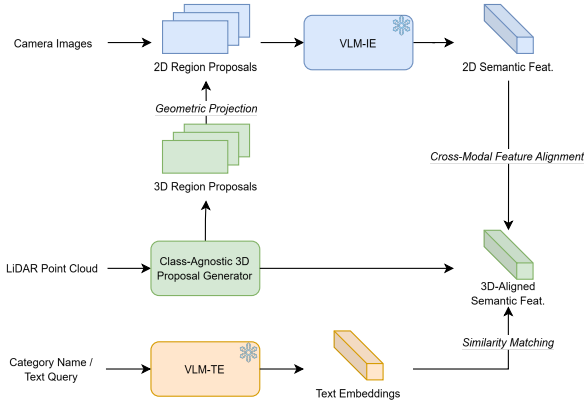


Figure 2.2: Representative geometry-driven 3D OV-OD architecture. A class-agnostic 3D proposal generator identifies candidate objects purely from LiDAR geometry. Semantic information is introduced by projecting 3D proposals into the image plane and aligning visual and textual embeddings from a pretrained VLM, decoupling geometric localisation from semantic identification.

point clusters as the primary geometric cue to discover candidate objects. Once these class-agnostic 3D bounding boxes are generated from the moving geometry, semantic assignment is achieved by unprojecting dense 2D vision-language features directly onto the 3D points within the box. While this temporal geometry approach circumvents the need for human annotations, its reliance on scene flow as the primary discovery mechanism makes the detection of static, unannotated objects highly challenging.

In contrast with previous works, Open3DWorld [10] adopts an alternative formulation by removing the conventional pipeline of class-agnostic 3D proposal generation followed by 2D semantic assignment. Instead, the method formulates OV-OD directly in LiDAR BEV space, where semantic information is incorporated through text-conditioned feature alignment. This is achieved by fusing BEV features from the point cloud with frozen CLIP text embeddings via cross-attention. Then, localisation and semantic prediction are performed jointly by computing similarity scores between BEV features and textual queries, followed by standard 3D bounding box regression. Despite this architectural difference, the method remains geometry-driven, as object discovery originates directly from LiDAR geometry.

Recently, OV-GT3D [61] extends this paradigm by revisiting the proposal scoring mechanism in class-agnostic 3D localisation. Instead of ranking proposals using a foreground–background objectness classifier, which may penalise valid but unannotated objects, the method introduces an Intersection-over-Union (IoU)-based two-stage detection framework. In the first stage, proposal quality is estimated via a dedicated IoU prediction head that evaluates geometric overlap between regressed boxes and their assigned training targets, independent of semantic category information. This design reduces the tendency of the proposal generator to overfit to seen classes, but at the cost of increasing proposal redundancy. To compensate for this, a second-stage Region of Interest (ROI) discriminator is used. While this strategy improves proposal generalisation, it primarily addresses proposal bias rather than fundamentally improving box regression accuracy.

AV3DOD [78] preserves the geometry-driven architecture introduced in CoDA, while removing the reliance on a predefined class vocabulary through an automatic semantic expansion mechanism that enables the prediction of labels beyond a fixed category set.

Overall, in AD contexts, sensor noise or calibration drift can cause the 3D-to-2D projection to fail, highlighting the fundamental bottleneck of geometry-first architectures: the inability of the semantic branch to compensate for geometric misses. Furthermore, the failure of these methods upon small objects is another critical challenge.

2.3.2 Vision-Driven Object Discovery

Vision-driven methods treat 2D visual understanding as the primary source of object semantics and supervision. Rather than discovering objects solely from 3D geometry, these approaches assume that mature 2D OV-OD models can reliably identify object regions in images, including categories unseen during training. The core idea is to leverage this rich 2D semantic signal to guide 3D localisation, either by lifting 2D detections into 3D space or by conditioning 3D prediction on image-derived geometric cues.

Frustum-Constrained 3D Learning. A defining characteristic of this subfamily is its reliance on offline pseudo-label generation pipelines. Because generating 3D bounding boxes from large 2D foundation models involves computationally demanding operations, these pipelines are executed as an offline pre-processing stage.

In this paradigm, 2D detections are projected into 3D search volumes (frustums) defined by camera calibration, and depth ambiguity is resolved through spatial clustering or geometric fitting. The resulting pseudo-labels serve as supervision targets for the downstream 3D detector. Consequently, the quality of downstream 3D proposal generation remains constrained by the geometric accuracy and robustness of the offline pseudo-labelling pipeline (illustrated in Figure 2.3).

OV-3DET [43] is one of the earliest methods to instantiate this paradigm by jointly operating on camera images and LiDAR point clouds. Given 2D detections, the method projects bounding boxes into 3D frustums and applies density-based clustering within the enclosed point cloud to estimate object centroids and suppress background points. The resulting pseudo 3D boxes are used to train a class-agnostic 3D proposal generator. While effective in transferring semantic information from 2D to 3D, this design exposes several geometric limitations. In particular, reliance on point cloud centroids often leads to orientation errors when objects are partially occluded or sparsely sampled, as shifts in the estimated centre directly affect bounding box rotation. Moreover, as the vocabulary expands, LiDAR sparsity increasingly limits the ability to distinguish distant objects, making semantically distinct categories geometrically ambiguous without dense visual cues. As a result, the quality of the generated 3D pseudo-labels becomes a critical bottleneck, tightly coupling the final 3D detector performance to both the reliability of the 2D OV-OD model and the robustness of the lifting procedure.

Moreover, OpenSight [79] attempts to resolve the heuristic failures of spatial clustering by integrating temporal continuity and Language-Model priors into its offline pseudo-label generation

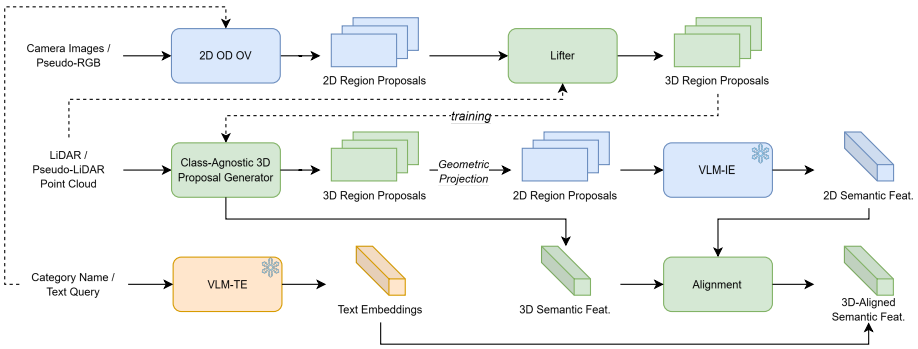


Figure 2.3: Representative vision-driven 3D OV-OD architecture based on frustum lifting. 2D open-vocabulary detections obtained from images are lifted into 3D frustum search volumes using camera geometry. These frustums constrain the generation of pseudo 3D proposals, which are used to supervise a class-agnostic 3D proposal generator. Semantic information is derived from image regions and textual queries via a pretrained VLM, while 3D localisation remains conditioned on the lifted geometric constraints.

pipeline. While it similarly uses 2D detections to isolate 3D frustums for point clustering, OpenSight addresses LiDAR sparsity at long ranges by projecting these initial bounding box estimates backward and forward across sequential frames to recover missed pseudo-labels. Furthermore, it introduces the use of categorical size priors (length, width, height) extracted via GPT-4 [49] to filter out geometrically implausible clusters and build a stable object bank for spatial augmentation prior to network training. While this combination of temporal tracking and LLM-derived priors significantly improves the quality of the supervisory data over early baseline methods, the reliance on predefined categorical size anchors inherently restricts the pipeline's ability to generalise to highly unusual or shape-dynamic objects that fall outside the LLM's statistical assumptions.

Find n' Propagate [15] also builds upon the frustum formulation but improves robustness by explicitly accounting for ambiguity in depth, scale, and orientation during proposal generation. Instead of generating a single 3D proposal per frustum, it densely samples multiple candidate boxes across these dimensions and selects the most plausible ones using a Greedy Box Oracle. This strategy reduces sensitivity to noisy geometric estimates. In addition, the method addresses a key bias towards nearby, dense objects by introducing a Remote Propagator that translates and rotates high-confidence samples to distant regions, together with a density simulator that mimics LiDAR sparsity through point dropout. These augmentations improve robustness to distance-related sparsity, representing a clear advancement over OV-3DET. However, semantic alignment continues to rely on VLM embeddings extracted from 2D regions and text queries and thus remains constrained by the performance of the underlying 2D detector.

ImOV3D [72] eliminates the need for annotated 3D data altogether and relies exclusively on large-scale 2D image datasets for supervision. RGB images are lifted into 3D using monocular depth estimation to generate pseudo point clouds, and 2D bounding boxes are projected into 3D frustums

and refined via clustering. To enforce physical plausibility, the method incorporates category-specific size priors obtained via GPT-4 [49] to filter implausible proposals. During inference, where only point clouds are available, ImOV3D bridges the modality gap by rendering 2D projections and using ControlNet [80] to hallucinate appearance, enabling VLM feature extraction. Unlike earlier methods, it fuses 3D geometric features and 2D semantic features into a joint representation prior to classification. This design allows ImOV3D to fully exploit the scale and diversity of 2D data, leading to performance gains over OV-3DET and CoDA. Nevertheless, the method remains sensitive to point cloud density and monocular depth noise, and the absence of camera extrinsics necessitates additional rotation correction to align the reconstructed geometry with a canonical ground plane.

OV-SCAN [11] extends this paradigm to outdoor autonomous driving scenarios. Similar to earlier methods, it relies on 2D object detections to initialise 3D frustums. However, instead of heuristic clustering or greedy proposal selection, OV-SCAN introduces an Adaptive 3D Box Search strategy to fit bounding boxes under sparsity. The method formulates 3D bounding box estimation as an explicit optimisation problem solved via Particle Swarm Optimisation (PSO). The objective jointly considers multiple geometric criteria, including maximising the number of enclosed LiDAR points, aligning box boundaries with visible structural edges, and anchoring the box surface to points closest to the sensor to reduce depth ambiguity.

Once these offline pseudo-labels are generated, they are used to supervise a 3D OV-OD architecture. Crucially, this downstream network predicts object spatial locations using category-dependent heatmaps, where each channel corresponds exclusively to a predefined base class. To enable OV recognition at inference, an alignment head maps the extracted 3D LiDAR features to the continuous 2D VLM embeddings captured during the offline phase.

Compared to centroid estimation [43], OV-SCAN yields more stable box fitting under sparse point clouds. However, the offline pipeline remains dependent on the quality of the initial 2D detections and exhibits degraded performance at long ranges, where insufficient LiDAR measurements limit the effectiveness of the optimisation. Furthermore, its reliance on predefined class anchor sizes constrains generalisation to categories with high intra-class shape variability, such as trailers or construction vehicles. Finally, because the downstream 3D architecture initiates object proposals directly from category-specific heatmaps, it may struggle to localise novel geometries that do not align with the predefined training classes.

Additionally, HQ-OV3D [37] further advances this family by introducing an explicit, learnt uncertainty modelling mechanism for proposal refinement. Rather than relying on heuristic sampling or optimisation alone, HQ-OV3D employs a diffusion geometric denoising module trained on annotated base classes. This module iteratively refines noisy 3D proposals by learning a probabilistic prior over translation, scale, and orientation, effectively transferring geometric knowledge from well-annotated categories to novel ones. This design substantially improves box quality under sparsity and occlusion, making the method effective for long-range detection in autonomous driving benchmarks. Nevertheless, performance remains influenced by the quality of the initial frustum proposals and the availability of sufficient geometric evidence for meaningful refinement.

Overall, frustum-constrained methods demonstrate strong potential by leveraging the semantic richness of mature 2D OV-OD models. Successive works have introduced increasingly sophisticated strategies for generating and refining 3D proposals under depth ambiguity and sparsity. Despite these advances, performance remains fundamentally bounded by the reliability of 2D detections and the heuristic nature of frustum lifting, motivating alternative formulations.

Explicit Geometry-Conditioned Box Learning. In contrast to frustum-based lifting, methods in this subfamily formulate 3D localisation as a direct regression problem conditioned on dense, image-derived geometric representations, rather than relying on heuristic proposal generation. As illustrated in Figure 2.4, these approaches lift visual information into a metric 3D space and predict object bounding boxes end-to-end, without intermediate pseudo-labels or frustum search.

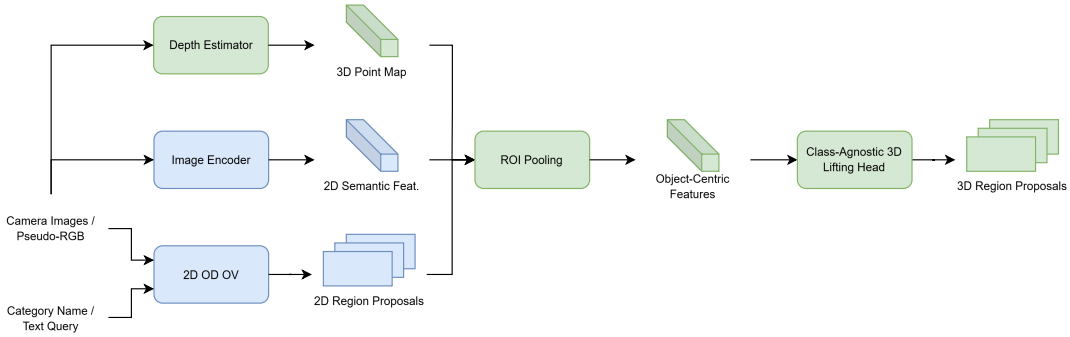


Figure 2.4: Representative vision-driven 3D OV-OD architecture based on explicit geometry-conditioned box learning. Image-derived geometric cues (e.g., monocular depth or learnt 3D features) and semantic representations from pretrained VLMs are aggregated into object-centric features via ROI pooling and directly regressed by a class-agnostic lifting head to predict metric 3D bounding boxes, without frustum-based proposal generation.

OVMono3D-LIFT [73] operates exclusively on RGB images during both training and inference and integrates a pretrained metric depth estimator to generate a dense 3D point map, alongside a self-supervised image encoder and a 2D OV-OD model for text-conditioned region proposals. These signals are fused through ROI pooling to produce object-centric features that explicitly encode geometric information, which are directly regressed by a class-agnostic 3D lifting head to predict metric 3D bounding boxes.

In contrast with ImOV3D [72] and other frustum-based approaches, where depth primarily serves to generate pseudo 3D supervision or constrain proposal search, OVMono3D-LIFT eliminates 3D proposal generation and pseudo-labelling entirely. This results in more stable orientation and scale estimation and avoids errors introduced by heuristic frustum sampling. However, the method remains fundamentally limited by monocular depth estimation accuracy and the availability of precise camera intrinsics, with depth errors propagating directly to 3D predictions. Performance degrades for small or distant objects, and the reliance on computationally heavy components such as DINOv2 and Grounding DINO increases inference cost. Additionally, the lack of explicit physical

priors can hinder generalisation to highly out-of-distribution categories with atypical shapes or sizes.

DetAny3D [77] further advances this paradigm by also operating purely on monocular RGB images, while placing stronger emphasis on exploiting large-scale 2D foundation models. Instead of explicitly constructing dense point maps, DetAny3D integrates representations from models such as SAM [26] and pretrained DINOv2 to capture rich semantic and geometric priors. These features are aggregated at the object level and lifted to metric 3D bounding boxes using a class-agnostic regression head.

A key contribution of DetAny3D is its stabilised knowledge transfer mechanism from 2D to 3D. Through a progressive geometry injection strategy, the method preserves the semantic richness of frozen 2D representations while gradually incorporating 3D spatial reasoning learnt from heterogeneous 3D supervision. This design allows it to more effectively leverage large-scale 2D priors while benefiting from 3D datasets. Consequently, DetAny3D achieves stronger performance than OVMono3D-LIFT, while still inheriting fundamental monocular limitations such as depth ambiguity for small or distant objects and increased computational cost due to heavy backbone models.

Overall, by directly regressing 3D bounding boxes from object-centric image features, these methods reduce sensitivity to frustum sampling and pseudo-label noise and better exploit large-scale 2D foundation models. Nevertheless, their reliance on monocular cues imposes fundamental limitations related to depth ambiguity and scale uncertainty.

2.3.3 Orthogonal Approaches

Beyond geometry-driven and vision-driven object discovery paradigms, a small number of methods explore orthogonal formulations for 3D OV-OD.

OV-Uni3DETR [68] proposes a unified multi-modal framework that enables bi-directional knowledge transfer, allowing large-scale 2D datasets to provide semantic supervision for 3D detection while 3D geometry constrains 2D learning. A switched-modality training strategy further enforces robustness under missing modalities, addressing the scarcity of annotated 3D data. While this enables strong generalisation across indoor and outdoor settings, performance remains coupled to the quality of pretrained 2D detectors and is challenged by extreme point cloud sparsity in outdoor scenarios.

OVM3D-Det [24] uses only RGB images. Object regions are first identified via Grounded-SAM [55] using arbitrary text prompts, while a monocular depth estimator predicts metric depth. Segmentation masks and depth maps are combined to back-project pixels into object-centric pseudo-LiDAR point clouds. To improve geometric plausibility, the method estimates the ground plane, derives object orientation via Principal Component Analysis (PCA), and incorporates GPT-4 [49] priors on object dimensions to filter implausible boxes. The resulting refined pseudo-labels are used to train a detector.

While this pipeline enables detection of novel categories without 3D annotations, its applicability to autonomous driving is limited by the reliability of monocular depth estimation. Depth errors

increase rapidly with distance, leading to poor far-range performance, and sensitivity to lighting and weather further affects robustness. This is reflected in a significant gap to oracle models trained with ground-truth 3D labels (e.g., 8.2% vs. 21.6% AP on novel classes in KITTI [17]), highlighting the limitations of pseudo-LiDAR supervision in safety-critical driving scenarios.

Additionally, TriCLIP-3D [31] directly targets challenges central to autonomous driving, particularly modality imbalance under outdoor sparsity. It introduces a cross-modal Mixture-of-Experts (MoE) architecture that dynamically routes information between geometric (LiDAR), semantic (image), and fused experts. This adaptive fusion allows the model to rely more heavily on visual cues when LiDAR points are sparse, which is common for distant or small objects, such as pedestrians and cyclists, while still exploiting geometry when available. This flexibility improves robustness and generalisation to novel classes in outdoor driving environments.

OP3Det [40] is a 3D class-agnostic detector. During training, class-agnostic 2D proposals generated by SAM are projected into 3D using multi-scale point sampling, and a cross-modal MoE dynamically balances geometric and appearance features. Trained with a binary objectness objective and without predefined vocabularies or text prompts, OP3Det achieves strong generalisation across datasets and environments, including outdoor driving scenarios. Its emphasis on objectness reduces dependence on category bias, though performance remains influenced by point cloud sparsity.

2.4 Datasets and Evaluation Metrics in Open-Vocabulary Object Detection

As perception models evolve towards open-vocabulary understanding, the requirements for both the datasets they are evaluated on and the metrics used to score them have shifted. This section reviews the most relevant datasets for outdoor AD and corresponding evaluation metrics.

2.4.1 Autonomous Driving Datasets and the Holdout Paradigm

To understand the current dataset landscape, Table 2.1 compares widely used closed-set AD datasets with the OpenAD [70] benchmark in terms of sensing modalities, annotation types, semantic coverage, and environmental diversity.

From Table 2.1, it is evident that while all listed datasets provide RGB data and 2D bounding box annotations, rich 3D support and broad semantic vocabularies are rare. Standard benchmarks like KITTI [17] and Waymo [60] rely on rigid, closed-set taxonomies with fewer than 10 classes. This significantly limits their direct use for OV-OD, where the ability to recognise rare or atypical objects is essential.

To bypass this limitation, recent efforts like the OpenAD [70] benchmark attempt to unify multiple datasets (e.g., nuScenes, Waymo, KITTI) into a single cross-dataset framework, using Multimodal Large Language Models (MLLMs) to automatically annotate over 200 distinct classes, as illustrated in Figure 2.5.

Table 2.1: Comparison of Standard AD Datasets and Benchmarks. Adverse conditions indicate the number of challenging factors represented (e.g., night, rain, fog). Adapted from Xia et al. [36, 70].

Dataset/Benchmark	RGB	LiDAR	2D BB	3D BB	Size	Classes	Adv. Cond.
nuScenes [6]	✓	✓	✓	✓	40K	23	3
CityScapes [12]	✓	✗	✓	✓	25K	30	1
BDD100K [75]	✓	✗	✓	✗	12M	40	5
Waymo [60]	✓	✓	✓	✓	230K	4	2
KITTI [17]	✓	✓	✓	✓	41K	8	1
Argoverse 2 [69]	✓	✓	✓	✓	6M	30	2
ONCE [46]	✓	✓	✓	✓	1M	5	3
OpenAD [70]	✓	✓	✓	✓	2K*	206	3

*OpenAD's size refers to 2,000 scenarios selected from other datasets.

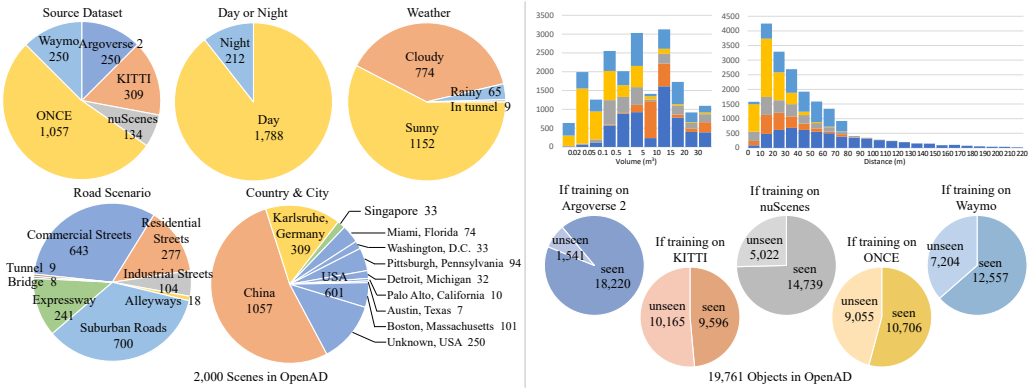


Figure 2.5: Statistical overview of the OpenAD benchmark, demonstrating cross-dataset semantic expansion. Source: Extracted from Xia et al. [70].

While cross-dataset benchmarks like OpenAD are valuable for testing domain generalisation, this paradigm inherently conflates domain shift with open-vocabulary generalisation. To evaluate the ability to recognise unseen categories, the literature commonly partitions single datasets (e.g., nuScenes) into base and novel classes [7, 43]. However, a methodological ambiguity persists in standard implementations: while ground-truth annotations for novel classes are removed from supervision, prevailing pipelines frequently expose the model to novel classes during training via offline pseudo-labels. Therefore, an explicit holdout protocol is required where all direct and indirect category-level supervision is absent during training by removing target classes from both ground-truth annotations and pseudo-label generation. Crucially, the underlying sensor returns (LiDAR points and camera pixels) corresponding to novel objects remain physically present in the driving scenes and are treated as background by the detector. This protocol tests whether the 3D detector can discover and localise unlabelled object geometries without category supervision during training, while acknowledging that upstream vision-language components (used for pseudo-labelling and text prompt embeddings) have encountered these semantic concepts during large-scale pre-training.

2.4.2 Evaluation Metrics

Standard evaluation protocols designed for closed-set object detection are not directly transferable to the OV paradigm. Traditional metrics like Mean Average Precision (mAP) implicitly assume that every test instance belongs to a known training class, thereby penalising correct predictions of unknown objects as false positives. As highlighted in recent OV-OD literature, this leads to misleading conclusions about a model's true generalisation and open-ended reasoning capabilities [5, 13, 70]. In this section, we review the principal evaluation metrics adopted by state-of-the-art OV detection methods.

Seen vs. Novel Class Detection (mAP). The gold standard for evaluating true open-vocabulary generalisation is to report mAP separately for base (seen) and novel (unseen) categories [7, 42]. This enables a direct analysis of how well a model preserves performance on known classes while generalising beyond the training vocabulary, ensuring the model has not simply suffered from catastrophic forgetting.

Emerging Open-Vocabulary Metrics. As the field of 3D OV-OD matures, recent benchmark proposals are attempting to capture the nuances of open-vocabulary perception beyond standard mAP. The OpenAD benchmark [70], in particular, introduces a suite of supplementary metrics to address these challenges. To evaluate the detector's capability to localise unannotated obstacles independent of classification accuracy, it reports the Average Recall (AR) of novel objects. Furthermore, it uses the Harmonic Mean of base and novel Average Precision (AP) to penalise models that over-predict novel objects at the expense of base categories. To address the restrictive nature of exact label matching, OpenAD incorporates Semantic Similarity-Aware Matching, where predictions are considered correct if their VLM-computed semantic embeddings exceed a predefined similarity

threshold with the ground truth. Finally, it tracks Absolute Open-Set Error (A-OSE) to quantify the specific risk of a network confidently misclassifying an unknown object into a known category.

2.5 Discussion: The Domain Gap and Baseline Selection

The literature reviewed in this chapter presents a trade-off between semantic versatility and geometric reliability. While geometry-driven methods preserve 3D spatial fidelity, they fail when sparse point clusters fail to trigger objectness priors. Conversely, vision-driven methods transfer rich 2D semantics but remain sensitive to the heuristic clustering errors inherent to frustum lifting.

Methodological Decisions Derived from the Literature. To directly address these bottlenecks, the experimental design of this thesis is grounded in concrete decisions adapted from three pivotal works:

1. **OV-SCAN [11]:** Selected as our primary benchmark baseline. OV-SCAN demonstrates the state-of-the-art capability of combining offline pseudo-labelling with continuous feature distillation for outdoor 3D open-vocabulary detection. We adopt its two-stage framework, but directly target its unconstrained global clustering and its category-dependent proposal heatmaps for improvement.

2. **TransFusion-L [3]:** Provides the foundational query-based LiDAR architecture inherited by OV-SCAN. Analysing its proposal generation mechanism reveals a core bottleneck for open-vocabulary generalisation: its reliance on category-specific heatmaps to initialise 3D object queries. This directly motivates our central architectural hypothesis: collapsing these multi-channel heatmaps into a single class-agnostic objectness channel to evaluate whether proposal discovery can function without predefined semantic routing.

3. **CLIP [53]:** Establishes the shared vision-language latent space used across our work. Instead of discrete classification heads, we leverage CLIP image embeddings as continuous supervision targets during training and match the resulting 3D latent features against CLIP text embeddings for zero-shot inference, allowing us to isolate geometric localisation from cross-modal semantic alignment.

Building upon these foundations, Chapter 3 details the mathematical formulation and reproduction of the baseline pipeline, and presents our proposed local pseudo-labelling and architectural modifications.

Chapter 3

Methodology

Building upon the limitations identified in Chapter 2, this methodology aims to explore the structural constraints imposed by heuristic spatial clustering and closed-set semantic heatmaps.

The chapter is structured as follows: Section 3.1 details the mathematical formulation of the selected vision-driven baseline, the OV-SCAN Baseline, while Section 3.2 documents the reproduction of its closed-source pseudo-label generation mechanism, the Global Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) Pipeline. Section 3.3 identifies the specific spatial and semantic limitations of this baseline, motivating the two core contributions of this work: an improved Localised Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Pipeline (Section 3.4), and the proposed OV-D3D architecture, which enforces single-channel proposal discovery (Section 3.5). Finally, Section 3.6 establishes the experimental configuration, detailing the zero-shot evaluation protocol used to validate generalisation to novel geometries.

3.1 The Baseline Architecture: OV-SCAN

The OV-SCAN Baseline [11] belongs to the Vision-Driven Object Discovery family of the 3D Open-Vocabulary Object Detection (OV-OD) paradigm, as detailed in Section 2.3.2. As illustrated in Figure 3.1, the architecture comprises two primary modules: an offline label generation pipeline (top), called Semantically-Consistent Novel-Object Discovery (SC-NOD), and the downstream 3D detection model (bottom). The label generation pipeline is responsible for creating a database of cross-modal proposals by pairing 2D detections with 3D point clouds, which subsequently serves as pseudo-ground truth to supervise the training of the 3D detector.

3.1.1 Label Generation Pipeline

In the 2D domain, Grounding DINO [39] is prompted with a set of novel classes to extract 2D bounding box proposals across K multi-view images. Simultaneously in the 3D domain, Patchwork++ [30] is used to filter ground points, and the HDBSCAN [44] algorithm is then used to cluster the remaining points.

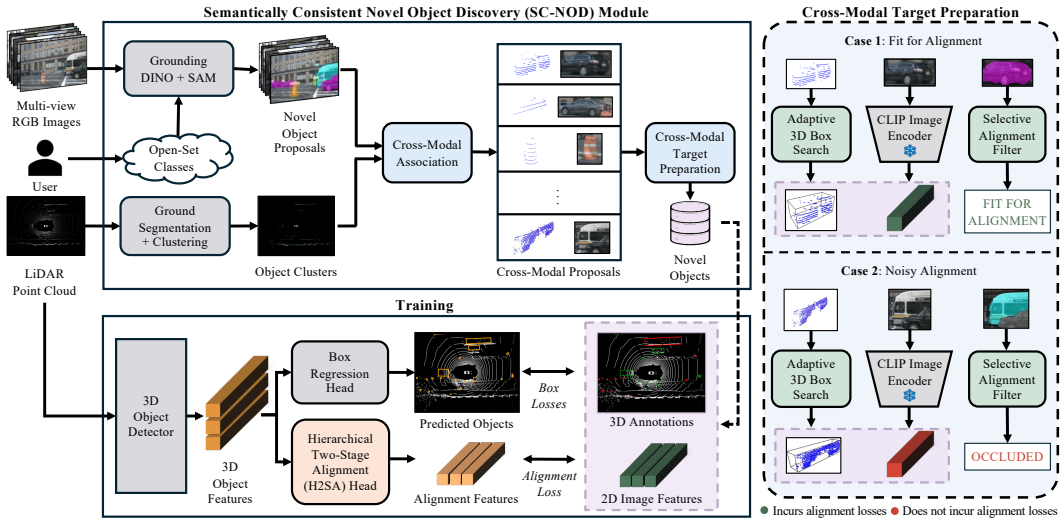


Figure 3.1: Overview of the OV-SCAN architecture. The system is decoupled into an offline pseudo-label generation pipeline (top) and a downstream 3D detection network (bottom). The offline pipeline generates cross-modal proposals by clustering LiDAR points and aligning them with 2D visual prompts, followed by an Adaptive 3D Box Search. The downstream 3D detector builds upon a grid-based backbone, replacing the standard classification head with an Hierarchical Two-Stage Alignment (H2SA) head. The H2SA head leverages Cross-Modal Distillation Blocks (CMDBs) to progressively upscale 3D geometric features into the Vision-Language Model (VLM) latent space. Source: Extracted from Chow et al. [11].

To establish cross-modal associations, each 2D bounding box is projected into 3D space to form a frustum. The pipeline pairs a 2D bounding box with a 3D cluster if they exhibit strong spatial correspondence, i.e., if the centroid of the 3D cluster falls within a predefined perpendicular distance threshold from the central ray of the projected 2D frustum.

Subsequently, for each cross-modal proposal comprising a 2D bounding box b_{img} and a set of 3D object points P_{obj}, the goal is to determine the optimal 3D bounding box parameters $\theta = (x, y, z, l, w, h, r_y)$ that best enclose the object. This is formulated as a continuous and non-differentiable optimisation problem. The search space is bounded by constraining the dimensions (l, w, h) to a predefined set of class-specific anchors $[A_{\min}, A_{\max}]$ and restricting the yaw angle $r_y \in [0, \pi]$. The optimisation seeks to minimise a total cost function J, defined as:

$$\min_{\theta} \quad J(\theta, P_{\text{obj}}, \mathbf{e}) = J_{\text{3D}}(\theta, P_{\text{obj}}, \mathbf{e}) + J_{\text{2D}}(\theta, b_{\text{img}}) \tag{3.1}$$

where $\mathbf{e}$ denotes the spatial coordinates of the ego-vehicle (sensor origin).

The 3D geometric cost, J_{3D}, is a weighted sum of three distinct spatial priors controlled by hyperparameters λ_1, λ_2, and λ_3, as follows:

$$J_{\text{3D}}(\theta, P_{\text{obj}}, \mathbf{e}) = \lambda_1 \cdot J_{\text{density}}(\theta, P_{\text{obj}}) + \lambda_2 \cdot J_{\text{l-shape}}(\theta, P_{\text{obj}}) + \lambda_3 \cdot J_{\text{surface}}(\theta, \mathbf{e}) \tag{3.2}$$

These represent specific geometric constraints applied to the bounding box:

- **Point Density Prior (J_{density}):** This component encourages the bounding box to encapsulate as much of the point cloud cluster as possible. It is defined as the negative ratio of interior points to total points:

$$J_{\text{density}}(\theta, P_{\text{obj}}) = -\frac{G_{\text{inside}}(\theta, P_{\text{obj}})}{|P_{\text{obj}}|} \tag{3.3}$$

  where G_{inside} counts the number of points falling within the 3D spatial volume $B_{\text{3D}}(\theta)$:

$$G_{\text{inside}}(\theta, P_{\text{obj}}) = \sum_{i=1}^{|P_{\text{obj}}|} \mathbb{1}\left(p_i \in B_{\text{3D}}(\theta)\right) \tag{3.4}$$

  with $p_i \in \mathbb{R}^3$ as the coordinate vector of the i-th point.

- **L-Shape Geometric Prior ($J_{\text{l-shape}}$):** Because LiDAR sensors typically only capture the surfaces of an object facing the sensor (forming an "L" shape for vehicles), this component forces the bounding box walls to align tightly with these visible points. It computes the mean orthogonal distance from all interior points to the nearest visible edge:

$$J_{\text{l-shape}} = \frac{1}{G_{\text{inside}}(\theta, P_{\text{obj}})} \cdot \sum_{i=1}^{|P_{\text{obj}}|} \min_{E \in \{E_1, E_2\}} d(p_i, E) \cdot \mathbb{1}\left(p_i \in B_{\text{3D}}(\theta)\right) \tag{3.5}$$

  where E_1 and E_2 represent the two edges of the bounding box closest to the sensor, and $d(\mathbf{p}_i, E)$ is the orthogonal distance from a point to an edge.

- **Surface Proximity Regulariser (J_{surface}):** This component regularises the depth of the bounding box relative to the sensor position $\mathbf{e}$:

$$J_{\text{surface}}(\theta, \mathbf{e}) = -\min\left(\left\|(\theta_x, \theta_y) - (\mathbf{e}_x, \mathbf{e}_y)\right\|_2, C_{\text{surface}}\right) \tag{3.6}$$

  where C_{surface} is a capping constant used to push the bounding box away from the ego reference point.

Finally, the cross-modal alignment cost J_{2D} aims at ensuring consistency between the 3D prediction and the original 2D visual proposal:

$$J_{\text{2D}}(\theta, b_{\text{img}}) = \gamma \cdot \left(1 - \text{IoU}_{\text{2D}}(P_{\text{3D}\to\text{2D}}(\theta), b_{\text{img}})\right) \tag{3.7}$$

where $P_{\text{3D}\to\text{2D}}(\theta)$ denotes the projection of the 3D bounding box corners onto the 2D camera image plane, IoU_{2D} calculates the Intersection-over-Union (IoU) with the original 2D detection b_{img}, and γ is a scaling weight.

Because the resulting objective function relies on discrete point counting and hard geometric boundaries, it is non-differentiable. Therefore, it is optimised using Particle Swarm Optimisation (PSO) [14].

The initial association step may map a single 2D proposal to multiple candidate 3D bounding boxes. To resolve this ambiguity, the pipeline retains only the 3D bounding box that achieved the lowest cost J during the PSO step. Furthermore, to eliminate redundant proposals of the same object captured across overlapping camera views, Non-Maximum Suppression (NMS) is applied in the 3D space.

Moreover, to estimate object velocities and filter noisy annotations, the authors used multi-object tracking: ImmortalTracker [66].

While all proposals retained from the previous step contribute to the regression loss, the authors demonstrated that restricting the alignment loss to only high-quality pairs during training yielded a higher final mAP. To achieve this, they excluded samples falling below a predefined resolution threshold. Furthermore, to mitigate occlusion corruption, they used Segment Anything Model (SAM) [26] to generate an instance mask for each 2D proposal, keeping only the cases where the object visibility exceeded class-specific thresholds.

3.1.2 OV-SCAN architecture

3.1.2.1 Base Architecture: TransFusion-L

OV-SCAN builds upon TransFusion-L [3], a state-of-the-art 3D object detector originally designed for closed-set detection. The architecture fundamentally operates by transforming sparse 3D geometry into a dense 2D representation for object discovery, followed by a transformer-based decoder.

Feature Extraction and BEV Projection. First, the raw LiDAR point cloud is voxelised, grouping the 3D points into a regular grid. A 3D Sparse Convolutional Neural Network, typically implemented as a Sparse ResNet backbone [71, 74], extracts geometric features while sequentially downsampling the spatial resolution by a factor of 8. To optimise computational efficiency, these sparse 3D features are collapsed along the Z-axis, flattening the data into a 2D Bird's Eye View (BEV) map. Standard 2D convolutional blocks then process this map to produce a rich, dense spatial feature representation, denoted as $\mathscr{F}_{BEV} \in \mathbb{R}^{C \times W \times H}$, where C is the feature dimension, and W and H are the spatial dimensions of the grid.

Heatmap Prediction and Query Initialisation. Following the standard TransFusion-L [3] detection mechanism, object discovery is driven by a dense heatmap. The network processes $\mathscr{F}_{BEV}$ to predict a multi-class heatmap $\mathscr{H} \in \mathbb{R}^{K \times W \times H}$, where K represents the number of predefined semantic categories. This heatmap estimates the probability of an object's centre residing at a specific spatial location for a given class. TransFusion-L then extracts the top N local maximums (e.g., $N = 200$ proposals) from this heatmap to serve as initial object candidates.

The initial object queries for the transformer decoder are subsequently formed by extracting the geometric features from $\mathscr{F}_{BEV}$ at these N peak locations and summing them with class-aware semantic embeddings derived directly from the heatmap's category channel.

Transformer Decoder and Regression. These initialised queries interact with the dense BEV feature map $\mathscr{F}_{BEV}$ via a Transformer Decoder. Through stacked self-attention and cross-attention mechanisms, the network aggregates surrounding contextual features, refining the queries into enriched 128-dimensional object embeddings. Finally, a set of feed-forward Multi-Layer Perceptrons (MLPs) decodes these embeddings to regress the final 3D bounding box parameters (centre coordinates, dimensions, rotation, and velocity).

3.1.2.2 Ground Truth Augmentation

To deal with the extreme class imbalance and sparsity inherent in autonomous driving datasets, state-of-the-art 3D detectors rely heavily on ground truth augmentation. OV-SCAN [11] extends this paradigm to support cross-modal alignment. Then, during training of the method, the augmentation was enabled for 15 epochs, but then disabled for the last 5 epochs. The process has two main stages:

Database Creation (Offline). For every pseudo-label generated by the SC-NOD pipeline, the system extracts the physical 3D LiDAR points inside the bounding box and saves them.

Online Scene Augmentation. During the training phase, the sampler class dynamically enriches the current scene to ensure the network sees a balanced distribution of objects.

3.1.2.3 OV-SCAN Modifications for Open-Vocabulary Detection

To bridge the gap between 3D geometry and 2D semantics, OV-SCAN replaces the standard classification head of TransFusion-L [3] with an H2SA head. The objective of the H2SA head is to upscale the 128-dimensional 3D geometric features into a 1024-dimensional latent space that aligns with the pre-trained CLIP [53] vision-language model. Specifically, this process operates as follows:

1. **Auxiliary Heatmap Classification**: The standard TransFusion-L heatmap is retained and trained to predict 10 predefined base classes (e.g., “car”, “pedestrian”). Once a 3D geometric query is extracted from the heatmap, its predicted base class is used to retrieve a pre-computed CLIP text embedding. This text embedding functions as a semantic anchor for the subsequent lifting process.

2. **Cross-Modal Distillation Blocks (CMDB)**: The H2SA head progressively upscales the 128-dimensional 3D geometric feature to 1024 dimensions across three sequential stages ($128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$). Within each CMDB stage, Multi-Head Cross-Attention is applied. Crucially, a transformed version of the retrieved base-class text embedding serves as the Query, while the 3D object features act as the Keys and Values. This mechanism forces the geometric upscaling process to be guided by closed-set language priors.

To supervise this semantic alignment during training, the final 1024-dimensional predicted feature is directly compared against the 1024-dimensional CLIP image embedding, extracted offline from the corresponding 2D visual crop, using a Cosine Embedding Loss.

Finally, at inference time, the network computes the dot product between the predicted 1024-dimensional alignment feature and a dictionary of pre-computed CLIP text embeddings for Open-Vocabulary (OV). The semantic label corresponding to the highest cosine similarity score is dynamically assigned to the 3D bounding box.

3.2 Baseline Reproduction and Implementation

To establish a solid experimental control, we reproduced the baseline system from the available resources. The original authors publicly released the 3D detector architecture and a pre-computed database of cross-modal proposals, but the offline pseudo-label generation pipeline remained entirely closed-source. Consequently, our reproduction effort fell into two distinct parts: adapting the released detector codebase and reverse-engineering the data generation pipeline from the paper.

For the downstream detector, we integrated the public architecture directly, adapting the training configuration to incorporate gradient accumulation to achieve the target effective batch size of 32 on our hardware without out-of-memory errors. The author-provided reference pseudo-labels were retained as a benchmark control to validate our detector implementation independently of data generation noise.

In contrast, the offline pseudo-label generation pipeline had to be reconstructed from scratch. While the published manuscript provided general architectural principles and occasional parameter hints (such as specific cost weights or iteration limits), critical operational specifics were omitted. We therefore had to determine missing configurations through systematic benchmarking and practical engineering decisions. Specifically, the prompt structure for Grounding DINO was empirically selected after evaluating concatenated versus per-class queries; the global HDBSCAN parameters had to be tuned across grid searches; and the cross-modal association logic was adapted to retain only the single best-matching cluster ($N = 1$) because the cost grows linearly with the number of matched clusters. Furthermore, while the paper specified PSO over 3,000 iterations, we observed stable convergence at 500 iterations, and introduced absorbing boundary conditions, velocity clamping, and modulo arithmetic for yaw to prevent the swarm from oscillating or getting trapped at boundary limits. Finally, temporal post-processing via ImmortalTracker was integrated by enforcing a birth threshold and discarding hallucinated interpolations. All parameters identified or tuned through these benchmarks are detailed in Appendix C.

2D Proposal Generation. The first phase consists of generating 2D visual proposals using Grounding DINO [39]. The used prompt was not specified, so, in order to reproduce this step, we empirically evaluated multiple prompt engineering strategies, ranging from single-class prompts to concatenated vocabularies, against the authors' reference database. The final implemented pipeline uses a concatenated string of the standard 10 nuScenes classes, as it presented the best trade-off

between the computational time and recall when evaluated against the OV-SCAN Reference Labels. The full evaluation of these strategies is detailed in Section 4.4.1.

Point Cloud Pre-processing and Clustering. In our implementation, Patchwork++ [30] was applied using its default configuration for the nuScenes dataset. However, prior to the ground removal phase, we implemented a default spatial cropping of the point cloud to a range of $X, Y \in [-51.2, 51.2]$ metres and $Z \in [-5.0, 3.0]$ metres.

Following cropping and ground removal, the points are processed by HDBSCAN [44]. Because this algorithm is highly sensitive to the sparsity variations inherent in LiDAR data, several experiments were conducted. The optimal configuration was a minimum cluster size of 5 points, a minimum samples parameter of 3, and a cluster selection epsilon of 0.8.

Cross-Modal Association. Our implementation bounds the 3D viewing frustum to a depth range of 2.0 to 60.0 metres from the camera sensor.

Our implementation projects all points of a candidate cluster onto the camera image plane. For the subset of points that fall inside the 2D bounding box boundaries, we calculate their perpendicular distance to the central ray of the frustum. A cluster is deemed a valid match only if its minimum point-to-ray distance is below a threshold ($\tau_{\text{match}} = 0.5$ metres).

Furthermore, dense scenes frequently result in a single 2D frustum encompassing multiple valid 3D clusters. While the original paper implies passing all valid clusters to the optimisation phase, our empirical ablations demonstrated that restricting the pipeline to the single highest-confidence candidate cluster ($N_{\text{matched}} = 1$) maintained downstream performance while significantly reducing computational overhead.

This cluster then forms a cross-modal proposal and undergoes the continuous PSO bounding box generation.

Adaptive 3D Box Search Implementation. To optimise the non-differentiable cost function J defined in Section 3.1, we implemented a PSO algorithm configured with $N_{\text{swarm}} = 50$ particles and executed over 500 iterations. While the original authors specified 3000 iterations, empirical observations indicated that the swarm achieved stable convergence at 500 iterations. Finishing the search earlier not only improved computational efficiency but also mitigated the risk of the swarm collapsing into implausible local minima.

To initialise the swarm, we follow the authors' strategy: half of the swarm's spatial coordinates (x, y, z) are initialised at the cluster's geometric mean, while the remaining half are initialised at the specific cluster point closest to the central camera ray. Gaussian noise ($C_{\text{noise}} = 0.1$) is subsequently injected to ensure sufficient spatial variance. To replicate the optimisation dynamics of OV-SCAN, the inertia weight (w) is dynamically decayed using a cosine annealing schedule from 10.0 to 0.1, following the authors' exact specified configuration.

While the theoretical bounds of the search space are defined in Section 3.1, ensuring the swarm actually converges to a valid solution requires some constraints. We introduced three

critical adaptations to the baseline formulation, adhering to the optimisation principles defined by Engelbrecht [14]:

- **Continuous Yaw Space:** Although the theoretical yaw angle r_y is restricted to $[0, \pi)$, clipping the value during optimisation causes an artificial accumulation of particles at the boundaries. To allow rotational exploration across the bounds, we implemented modulo arithmetic, mapping any updated angle back to the valid domain via $r_y \leftarrow r_y \pmod{\pi}$. This ensures that negative rotations or over-rotations wrap smoothly without boundary trapping.
- **Velocity Clamping:** To control global exploration and prevent the search from exploding outward into zero-gradient regions, we explicitly clamp the maximum velocity $V_{\max,j}$ for each dimension j as a proportion of the domain range: $V_{\max,j} = \delta(x_{\max,j} - x_{\min,j})$. Following standard practices, we set $\delta = 0.5$.
- **Absorbing Boundaries:** If a particle attempts to breach the spatial limits defining the positional or dimensional bounds $[A_{\min}, A_{\max}]$, its position is immediately clamped to the boundary constraint, and its corresponding velocity component $v_{ij}(t)$ is zeroed. This absorbing boundary condition allows particles to slide along the constraints without violent oscillation.

The objective function J was evaluated using the authors' hyperparameters $\lambda_1 = 5.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, $\gamma = 3.0$, and $C_{\text{surface}} = 5.0$. Finally, the class-specific dimensional anchors were derived from OpenPCDet conventions and allowed a tolerance of $\pm 20\%$.

Post-Optimisation Filtering and NMS. To eliminate redundant predictions of the same physical object captured across different cameras, a global 3D NMS is applied to the entire scene. We use a 3D IoU threshold of 0.50. To ensure higher precision for rotated bounding boxes, the 3D IoU is calculated by projecting the 3D bounding boxes into a BEV and computing the precise polygon intersection area using the Sutherland-Hodgman clipping algorithm [62].

Selective Alignment Preparation. We followed the authors' approach by calculating the pixel mask occupancy ratio of each 2D proposal by using the SAM [26] ViT-H model. Then, we used their proposed visibility thresholds to filter out heavily occluded proposals. We assigned a visibility threshold of 0.50 to the `trailer` class, as it was missing in the original paper.

Temporal Tracking Integration. As the final step in the baseline label generation pipeline, the original architecture uses ImmortalTracker [66] to estimate object velocities and filter noisy annotations across sequential frames.

In our reproduction, we enforced a track-birth threshold, requiring a track to be consistently detected across at least two sequential frames before initialisation.

To integrate the tracker's trajectories back into our frame-level training database, we associate them via the Hungarian Algorithm. This matching logic yields two outcomes:

1. **Matched Proposals:** Tracked boxes that match an original cross-modal proposal inherit the tracker's estimated velocity vector and retain their CLIP embedding. Crucially, original proposals that fail to associate with any established temporal track are classified as transient noise and are discarded.

2. **Interpolated Proposals:** While the tracker can mathematically hallucinate bounding boxes in frames where the object is occluded, relying on these interpolated boxes injects geometric noise. Therefore, in the final implementation, purely interpolated proposals are discarded. The impact of this design choice is analysed in Section 4.4.2.

3.3 Limitations of the Baseline

The baseline relies on spatial heuristics and closed-set assumptions that limit its open-vocabulary generalisation. These limitations are detailed below.

3.3.1 Limitations of the Pseudo-Label Generator

Dependence on Anchor Boxes. The SC-NOD pipeline relies on predefined class anchors to bound the 3D box search. This creates two problems. First, if the 2D VLM misclassifies an object (e.g., due to occlusion), the pipeline applies the wrong dimensional prior, resulting in an incorrect 3D bounding box. Second, expanding the model to novel categories requires manually defining new anchor sizes, which contradicts the goal of OV-OD.

Post-Hoc Temporal Integration. The baseline incorporates temporal data as a post-processing step by applying ImmortalTracker [66] to the final 3D pseudo-labels. Because these 3D boxes are generated heuristically on a per-frame basis, they are inherently noisy, making downstream 3D tracking unreliable. In contrast, 2D object detections are significantly more mature and stable across sequential frames. By delaying temporal integration to the end of the pipeline, the system cannot leverage temporal history to aggregate geometry and resolve single-frame sparsity before the 3D bounding box optimisation occurs.

Global 3D Clustering. Applying HDBSCAN [44] globally across a scene fails to account for the radially degrading density of LiDAR point clouds. A configuration tuned to cluster sparse, distant points will over-segment dense, nearby objects. Conversely, tuning for the near field will under-segment distant objects. Because 2D detectors already isolate the object reliably in the image plane, clustering globally across the whole 3D scene introduces unnecessary segmentation errors that could otherwise be constrained by the available 2D visual priors.

3.3.2 Limitations of the Model[1]

The baseline training architecture introduces closed-set priors into the geometric learning process through two mechanisms.

Embedding Upscale Bias. The H2SA head uses predefined base-class text embeddings as queries in its cross-attention blocks. This forces the upscaling of 3D features to be guided by base-class semantics, potentially restricting the latent space's ability to represent novel objects.

TransFusion Heatmap and Category-Dependent Biases. As mentioned in Section 3.1.2.1, the baseline architecture relies on predicting a dense BEV heatmap, $\hat{S} \in \mathbb{R}^{K \times W \times H}$, where K corresponds to the number of base categories. Initial object queries are generated by extracting the top-N absolute peaks across all K channels. The architecture then explicitly injects class-aware embeddings into these queries by encoding the origin channel of the peak as a one-hot vector and adding it to the geometric features.

However, in an open-vocabulary context, this multi-class formulation introduces structural limitations. By having K distinct semantic channels in the heatmap, the base vocabulary is explicitly hardcoded into the geometric backbone. Instead of learning to detect universal objects, the network learns to filter for predefined shapes. Consequently, queries extracted for novel classes ($\mathscr{C}_n$) remain inherently biased towards the spatial distributions of the K base classes.

Furthermore, this architecture forces the geometric backbone to implicitly trust the 2D VLM classification due to the channel-specific loss formulation. If a single spatial location triggers a peak in multiple class channels simultaneously, the network spawns several proposals initialised with conflicting semantic priors (e.g., one query geometrically biased as a `Motorcycle` and another as a `Pedestrian`). For highly sparse or occluded objects where the offline VLM prediction is noisy or ambiguous, hardcoding this potentially erroneous label into the geometric extraction phase may corrupt the resulting query.

Ultimately, these limitations suggest that relying on intermediate closed-set classification may restrict pure object discovery in 3D OV-OD.

This thesis systematically addresses these limitations. Specifically, the issues of Global 3D Clustering and category-dependent biases are addressed by our improved label generation pipeline (Section 3.4) and the proposed OV-D3D architecture (Section 3.5). The limitations regarding Anchor Box Dependence and Post-Hoc Temporal Integration were also explored, and our experimental attempts to resolve them are detailed in Appendix E.

[1] Portions of the text in this section have been adapted from the workshop paper: Angélico, P., et al., "Decoupling Geometry and Semantics for Robust Open-Vocabulary 3D Object Detection" [1], accepted at the European Conference on Computer Vision (ECCV) 2026.

3.4 Improved Pseudo-Label Generation Pipeline

The development of our pseudo-label generation pipeline was motivated by the observation that applying hierarchical clustering globally across an entire outdoor LiDAR scan is ill-suited to radial point decay. In the baseline setup, running HDBSCAN across the full scene causes substantial over-segmentation of dense foreground objects and under-segmentation at range. While projecting 2D detections into 3D frustums is standard practice to establish visual bounds, we hypothesised that clustering exclusively within the isolated points of each individual frustum would be sufficient to extract the target geometry while reducing spatial heterogeneity. By restricting the clustering domain to the points inside the frustum, the density variance of the full driving scene is reduced. Because the target object occupies a bounded depth slice, we replaced the global HDBSCAN with a localised DBSCAN implementation ($\varepsilon = 0.8$, min_samples $= 4$; see Appendix C). Crucially, because this pipeline alters both the clustering domain (from global scene to local frustum points) and the algorithm (from HDBSCAN to DBSCAN), subsequent evaluation measures the aggregate benefit of this combined spatial and algorithmic intervention rather than isolating either factor independently. The effectiveness of this design is evaluated through geometric precision, recall, and F1-score against both nuScenes ground-truth annotations and the reference database at 3D IoU ≥ 0.5, followed by downstream 3D detection performance.

All subsequent optimisation stages from the baseline pipeline, including the PSO bounding box fitting, are retained without alteration.

Specifically, Figure 3.2 illustrates the step-by-step execution of this pipeline, from the initial 2D Grounding DINO detection through to the final 3D bounding box convergence.

3.5 The Proposed OV-D3D Architecture[2]

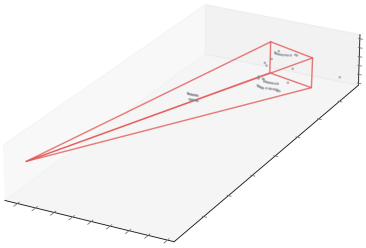
The design of the proposed OV-D3D architecture was motivated by the observation that prevailing query-based detectors condition 3D proposal generation on multi-channel categorical heatmaps. In an open-vocabulary setting, this mechanism binds initial spatial queries to predefined base classes, causing competing queries at shared locations and penalising unannotated geometries. We hypothesised that geometric proposal discovery fundamentally benefits from removing these intermediate closed-set classification priors. To isolate and test this core hypothesis, we collapsed the multi-class heatmap into a single class-agnostic objectness channel.

However, eliminating categorical channels deprives the network of category-specific scale priors and receptive-field shortcuts, which are crucial when distinguishing objects with extreme scale disparities (such as pedestrians versus barrier walls or construction vehicles) within the exact same spatial grid. To compensate for this loss of structural information, we introduced two complementary modules: a Gated Atrous Spatial Pyramid Pooling (ASPP) module to dynamically adapt the spatial receptive field across diverse geometries, and Harmonic Scale Embeddings to reintroduce

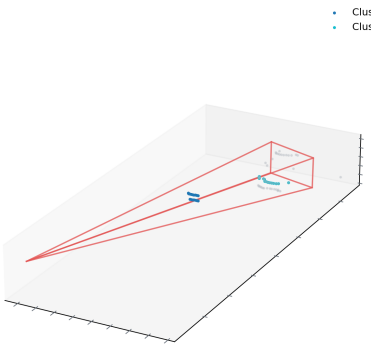
[2]Portions of the text in this section have been adapted from the workshop paper: Angélico, P., et al., "Decoupling Geometry and Semantics for Robust Open-Vocabulary 3D Object Detection" [1], accepted at the European Conference on Computer Vision (ECCV) 2026.



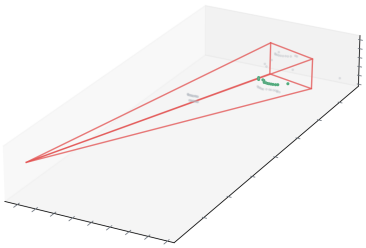
(a) Detected 2D object with Grounding DINO.



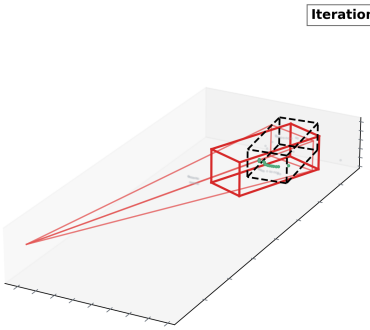
(b) Extracted 3D Frustum for the detected object.



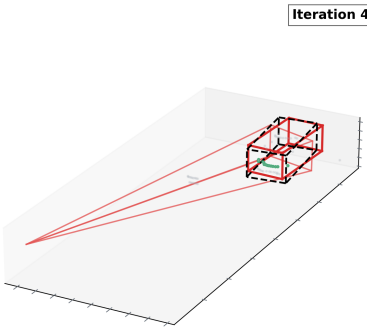
(c) Local DBSCAN Clustering inside the frustum.



(d) Target Cluster Selection based on centre-ray distance.



(e) PSO Initialisation (Iteration 0).



(f) PSO Convergence (Final Iteration).

Figure 3.2: Step-by-step visualisation of the proposed pseudo-label generation pipeline. **(a)** The target object is identified in 2D. **(b)** The 2D box is projected into a 3D frustum. **(c)** Local DBSCAN separates points into distinct geometric clusters. **(d)** The target cluster is isolated by selecting points closest to the camera's central ray. **(e)** PSO initialises a bounding box. **(f)** PSO converges, aligning the 3D bounding box.

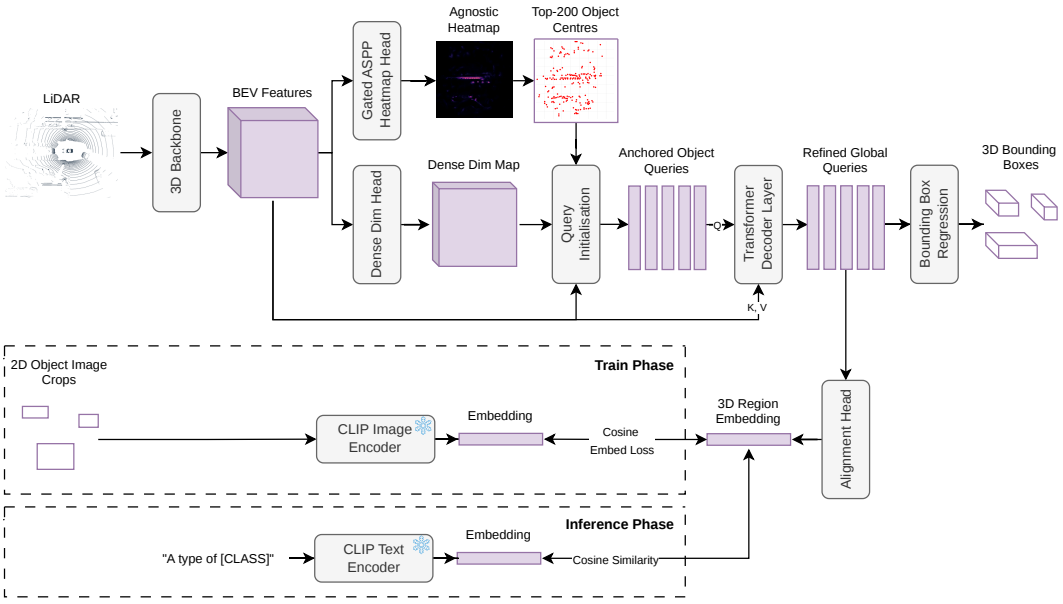


Figure 3.3: Overall architecture of OV-D3D. Raw LiDAR point clouds are processed into dense BEV features. Unlike previous methods, we enforce a 1-channel objectness heatmap to remove closed-set semantic priors from the query initialisation phase. Bounding box dimensional priors are densely predicted, encoded via Harmonic Scale Embeddings, and combined with gathered BEV features to initialise the Transformer Decoder. The refined global queries are processed by regression and FFN alignment heads. During training, the alignment head is supervised using offline CLIP image embeddings and during inference, it is compared against CLIP text embeddings.

continuous dimensional priors without restoring categorical labels. Crucially, these components do not form the central hypothesis itself. Instead, they serve as compensatory mechanisms designed to recover the spatial localisation and scale sensitivity lost when transitioning to a single-channel formulation. The validity of this design is evaluated through macro Mean Average Precision (mAP) on base classes against both the multi-channel baseline and a 1-channel control, alongside Class-Agnostic Average Recall (Mean Average Recall (mAR)) under holdout zero-shot conditions.

As illustrated in Figure 3.3, OV-D3D removes intermediate category classification from proposal generation. Once candidate boxes are initialised and refined, they are projected into the VLM latent space using a lightweight Feed Forward Network (FFN) alignment head, bypassing category-conditioned distillation.

3.5.1 Class-Agnostic Geometric Discovery

Gated ASPP Heatmap Head. To prevent the hardcoding of base vocabularies into the geometric backbone, we replace the standard multi-class heatmap with a 1-channel objectness map: $\hat{S} \in \mathbb{R}^{1\times W\times H}$. However, collapsing all categories into a single channel forces the network to simultaneously localise geometries of vastly different scales (e.g., pedestrians versus buses) within the

exact same spatial grid. Because a standard 3×3 convolution possesses a rigid receptive field, it inherently struggles to pinpoint accurate centres across this extreme scale variance.

To structurally adapt to this scale variance, we implement an ASPP architecture. The input BEV features are processed in parallel through four convolutional branches: a 1×1 convolution and three 3×3 convolutions with dilation rates of 1, 3, and 5. Simultaneously, a spatial gating generator applies a 3×3 convolution, followed by Batch Normalisation and a ReLU activation, before a final 1×1 projection to output a 4-channel spatial weight map. We apply a softmax over the channel dimension to generate pixel-wise competition weights. The four dilated feature maps are multiplied by their respective spatial weights and summed, allowing the network to adaptively fuse the optimal receptive fields for each specific location. This fused feature map is then projected into the final 1-channel objectness map, which is supervised via a Gaussian Focal Loss ($\mathscr{L}_{hm}$) following CenterPoint [74]. We apply 3×3 MaxPool non-maximum suppression to extract the top-N spatial indices corresponding to the highest objectness peaks.

Dense Dimension Prior and Query Initialisation. While localising objects in a 1-channel heatmap removes early semantic bias, it hinders bounding box estimation. Without semantic channels to act as size priors, extreme scale variance makes accurate dimension estimation harder. To address this, we introduce a parallel Dense Dimension Head, a convolutional block that densely predicts the 3D bounding box dimensions (l, w, h) in log-space for every location in the BEV grid, supervised via an L1 Loss ($\mathscr{L}_{bev_dim}$). This dense prediction serves as a structural prior to guide the downstream attention mechanism.

At the top-N extracted objectness peaks, we gather both the raw 128-D BEV features and the corresponding predicted log-space (l, w, h) scalars. Direct concatenation of low-dimensional scalars into high-dimensional feature spaces induces frequency bias and degrades Transformer dot-product attention [64]. Inspired by Dynamic Anchor Box formulations [38], we first pass the raw log-space 3D dimensions directly through a Harmonic Scale Embedding:

$$PE(s)_{2i} = \sin\left(\frac{s}{\tau^{2i/D}}\right), \quad PE(s)_{2i+1} = \cos\left(\frac{s}{\tau^{2i/D}}\right), \tag{3.8}$$

where D is the embedding dimension and the temperature is set to $\tau = 20.0$. To combine these modalities, the high-frequency positional embeddings are zero-padded to match the 128-D channel dimensionality of the backbone, processed by a 2-layer MLP, and added element-wise ($\oplus$) to the gathered BEV features, yielding the structurally stabilised Anchored Object Queries.

3.5.2 Vision-Language Alignment

The Anchored Object Queries interact with the dense LiDAR BEV features via cross-attention in the Transformer Decoder, producing context-aware Refined Global Queries. At this stage, the network splits into two branches.

The geometric pathway regresses the standard 3D bounding box parameters. The semantic pathway maps the 128-D geometric queries into the 1024-D CLIP latent space. To achieve this,

we instantiate a straightforward FFN Alignment Head: a linear projection to a 512-D hidden state, Layer Normalisation, GELU with dropout, and a final projection to the 1024-D CLIP space followed by Layer Normalisation.

The transition to an FFN rather than the baseline's complex H2SA head is a necessity of our formulation. As defined in Section 3.1, the baseline's H2SA head relies on predicting a base category from a multi-channel heatmap to retrieve a CLIP text embedding, which then serves as the Query for its cross-attention blocks. Because OV-D3D explicitly removes this categorical heatmap to eliminate closed-set biases, the architecture no longer produces the intermediate semantic predictions required to instantiate the H2SA module.

Instead, we formulate semantic alignment as a pure feature projection task. By using an FFN, the geometric features are mapped directly into the CLIP space without requiring intermediate closed-set text anchors. As will be evaluated in the ablation studies in Section 4.4, this formulation maintains competitive performance when ablated against more complex heads.

During the Training phase, 3D Region Embeddings that successfully match with valid 2D object crops are supervised via a Cosine Embedding Loss against their corresponding 1024-D CLIP image embeddings. During the Inference phase, the 3D Region Embeddings are directly compared against CLIP text embeddings generated using the single prompt template "`A type of [CLASS].`".

3.5.3 Query Decoding and Optimisation

During training, we use a class-agnostic Hungarian Assigner to formulate a bipartite matching between the N predicted queries and the pseudo-ground truth boxes, relying on a weighted sum of focal loss objectness cost, L1 distance, and 3D IoU costs. Once assigned, the bounding box regression is supervised via an L1 Loss ($\mathscr{L}_{box}$), and the query objectness confidence is supervised via a Sigmoid Focal Loss ($\mathscr{L}_{cls}$).

To supervise our newly introduced Dense Dimension Head, we formulate an auxiliary L1 loss ($\mathscr{L}_{bev_dim}$). While the head predicts dimensions densely across the entire BEV grid, the loss is masked such that it is only computed at the exact spatial indices corresponding to the ground-truth object centres. The total end-to-end training objective is defined as:

$$\mathscr{L}_{total} = \lambda_{hm}\mathscr{L}_{hm} + \lambda_{dim}\mathscr{L}_{bev_dim} + \lambda_{cls}\mathscr{L}_{cls} + \lambda_{box}\mathscr{L}_{box} + \lambda_{align}\mathscr{L}_{align}, \tag{3.9}$$

where $\mathscr{L}_{align}$ represents the Cosine Embedding Loss used to distil the CLIP image embeddings into our lightweight FFN alignment head. In our experiments, we empirically set the loss weighting coefficients to $\lambda_{hm} = 1.0$, $\lambda_{dim} = 1.0$, $\lambda_{cls} = 0.25$, $\lambda_{box} = 0.25$, and $\lambda_{align} = 1.0$.

3.6 Experimental Setup

To evaluate the proposed OV-D3D architecture and validate our reproduction of the baseline, this section details the datasets, the zero-shot holdout training protocol, the hardware configuration, and

the diagnostic metrics used to isolate geometric discovery from semantic alignment.

3.6.1 Datasets

This work uses the nuScenes [6] autonomous driving dataset, which serves as the primary benchmark for our baseline, OV-SCAN [11]. To validate our reproduction of the Global HDBSCAN Pipeline, we used the OV-SCAN Reference Labels, a database of cross-modal proposals provided by the original authors. Detailed distributions and file structures of this reference database are provided in Appendix B.

3.6.2 Subset Preserving Class Distribution

Training the OV-SCAN Baseline on the complete nuScenes dataset under the Full Benchmark Protocol requires approximately two days of continuous compute on available hardware, posing a practical bottleneck for iterative hyperparameter tuning. To enable efficient experimentation without distorting the dataset's natural long-tail class balance, we implemented a greedy selection algorithm to construct a subset selected to preserve the class distribution used in the full dataset. Rather than relying on naive random sampling, the algorithm iteratively evaluates candidate scenes and minimises the L1 distance between the subset's class distribution and the global target distribution of the full dataset. The mathematical formulation and empirical validation of this selection algorithm are detailed in Appendix B.2.

3.6.3 The Zero-Shot Holdout Protocol

As discussed in Section 2.4, many existing OV methodologies inadvertently leak semantic priors by using VLMs to generate pseudo-labels for novel classes during training. To evaluate open-vocabulary generalisation without training leakage, we establish the Zero-Shot Holdout Protocol, where explicit category supervision is removed for selected classes during the training phase.

The standard 10 annotated nuScenes categories are partitioned into two disjoint subsets. The training process is restricted entirely to seven base categories: `Car`, `Truck`, `Bus`, `Construction Vehicle`, `Pedestrian`, `Bicycle`, and `Traffic Cone`. The remaining three categories (`Motorcycle`, `Barrier`, and `Trailer`) are designated as the holdout novel classes ($\mathscr{C}_n$).

Because the detector is trained exclusively on offline pseudo-labels rather than human annotations, we prevent supervisory leakage by filtering out all generated pseudo-labels that exhibit a 3D IoU greater than 0.3 with any ground-truth holdout instance. Crucially, the sensor returns (LiDAR points and image pixels) corresponding to these novel objects remain physically present in the training scenes and are treated as background by the network. This protocol tests whether the detector can discover and localise unannotated geometries without explicit supervision. This does not imply, however, that the upstream pretrained 2D backbones have never encountered the visual concepts during their internet-scale pre-training.

3.6.4 Training Setup

The 3D detectors were trained end-to-end for 20 epochs using the AdamW optimiser with a weight decay of 0.01. The learning rate was modulated via a One-Cycle learning rate schedule, peaking at a maximum learning rate of 1×10^{-3}, with a division factor of 10 and momentum cycling between 0.805 and 0.9. Following standard practices for this architecture, ground-truth augmentation (Copy-Paste) was enabled for the first 15 epochs and explicitly disabled for the final 5 epochs to allow the network to adapt to natural scene distributions.

Regarding hardware constraints, the original OV-SCAN architecture specifies an effective batch size of 32. Replicating this directly on the available four NVIDIA V100 GPUs resulted in out-of-memory errors. To preserve the optimisation dynamics without exceeding hardware limits, we implemented Gradient Accumulation. Specifically, the training uses a per-GPU batch size of 4 paired with two gradient accumulation steps, yielding the target effective batch size of 32. Preliminary ablations (Section 4.1.2) confirm this configuration closely approximates the optimisation dynamics of the original design without performance degradation. An exhaustive list of all architectural hyperparameters, loss weights, and network configurations is provided in Appendix C.

3.6.5 Evaluation Metrics

Validating the pseudo-label generation pipeline through end-to-end model training introduces a computational bottleneck. To decouple the evaluation of the proposal generator from the 3D detector's training dynamics, we rely on intermediate auxiliary metrics:

- **2D Proposal Metrics:** We evaluate 2D bounding box recall at specific IoU thresholds against the OV-SCAN Reference Labels. Furthermore, to verify probabilistic consistency, we compute the Mean Absolute Error (MAE) between the confidence scores of our matched 2D detections and the reference labels.
- **3D Proposal Metrics:** To evaluate spatial accuracy, we compute precision, recall, and the F1-score of the Reproduced Global Labels and the Local Labels against both the OV-SCAN Reference Labels and the standard nuScenes ground truth.

Standard 3D Detection Metrics For downstream 3D detection evaluation, we report standard nuScenes metrics, primarily mAP and the nuScenes Detection Score (NDS). To provide a fine-grained analysis of bounding box regression quality across different architectural configurations, we additionally report True Positive metrics, specifically Mean Average Translation Error (mATE), Mean Average Scale Error (mASE), and Mean Average Orientation Error (mAOE).

Extended Evaluation Protocols To systematically evaluate the structural bottlenecks identified in the baseline (Section 3.3), we used the following evaluations:

- Class-Agnostic Localisation: Standard mAP penalises correct spatial localisations if the predicted semantic class is incorrect, conflating geometric failure with semantic failure. To isolate the pure discovery performance of the geometric backbone independently of its semantic classification, we map all predictions and ground truth annotations to a single universal "object" class and compute Class-Agnostic Average Recall. Following nuScenes conventions, this is primarily evaluated using a 2.0 m centre-distance matching threshold. Furthermore, to distinguish between macro-discovery and fine localisation accuracy, we report agnostic recall across distance thresholds of 0.5 m, 1.0 m, and 4.0 m.

- Environmental and Distance Splits: To reproduce the baseline's extended evaluation, the validation set is partitioned into environmental subsets (Day vs. Night, Dry vs. Wet) using free-text description keyword matching. Furthermore, detection performance is evaluated at three distance ranges from the ego-vehicle: Near (0–18 m), Medium (0–34 m), and Far (0–54 m).

Chapter 4

Results

This chapter presents the empirical evaluation of the proposed pseudo-label generation pipeline and the OV-D3D architecture. To enable rapid iterative ablations, some experiments in this chapter were conducted on a 30% subset of the nuScenes [6] dataset selected to preserve the global class distribution. The greedy selection algorithm used to construct this subset, alongside its validation metrics, are detailed in Appendix B.2.

To ensure unambiguous comparisons, the results in this chapter are evaluated across three distinct supervision conditions:

1. **OV-SCAN Reference Pseudo-Labels:** The pre-computed cross-modal labels provided by the authors of OV-SCAN [11], serving as an upper-bound benchmark control to validate detector implementations independently of data generation variations.

2. **Reproduced Global Labels:** Pseudo-labels generated by our reproduction of the baseline offline pipeline (Patchwork++ and Global HDBSCAN), establishing an internal baseline control.

3. **Proposed Local Labels:** Pseudo-labels generated by our frustum-constrained local pipeline (Local DBSCAN inside 2D Grounding DINO frustums).

4.1 Establishing the Baseline

To evaluate the proposed methodologies, the reproduction of the OV-SCAN [11] baseline was validated in isolated stages, beginning with the evaluation infrastructure, followed by the training mechanics, and concluding with the offline label generation pipeline.

4.1.1 Validating the Codebase and Evaluation Metrics

Before training any models, it was necessary to verify the correctness of our codebase and the accuracy of the reverse-engineered evaluation metrics (detailed in Section 3.6.5). To isolate the evaluation logic from any training variance, we executed our independent validation script using

the authors' officially released pre-trained weights. The results were then compared against the metrics published in the original OV-SCAN paper.

Table 4.1: Validation of the reverse-engineered evaluation script. The authors' pre-trained weights for the OV-SCAN Baseline are evaluated using our metric pipeline under the Full Benchmark Protocol and compared against the published results. The Δ column indicates the absolute deviation.

Metric	**Published Baseline**	**Pre-trained Weights**	Δ
	Weather and Lighting Conditions		
Day	31.1	31.1	+0.0
Night	23.4	23.3	−0.1
Dry	31.8	31.9	+0.1
Wet	25.6	25.6	+0.0
	Distance Bins		
Near (0–18 m)	48.1	48.2	+0.1
Mid (0–34 m)	35.6	35.8	+0.2
Far (0–54 m)	27.7	27.6	−0.1
	Overall		
mAP	31.1	31.0	−0.1
NDS	32.8	32.7	−0.1

As demonstrated in Table 4.1, our evaluation script replicates the published metrics with negligible deviation.

To verify the baseline generation of 2D proposals, Figure 4.1 provides a direct visual comparison against the OV-SCAN Reference Labels. While our pipeline recovers the spatial localisation of the objects, predicting nearly identical bounding boxes, we observe a divergence in global probabilistic confidence (explored quantitatively in Section 4.4.1). Because the original authors' exact Grounding DINO [39] textual prompts and temperature scaling parameters remain undisclosed, our reverse-engineered pipeline yields a different confidence distribution. Nevertheless, for highly salient objects, the raw detection scores often align closely (e.g., the motorcycle on the right registers a confidence of 0.74 in our reproduction against 0.75 in the baseline). Furthermore, the integration of Segment Anything Model (SAM) [26] instance masks (Figure 4.1, Left) highlights the persistent challenge of 2D occlusion, such as the bicycle overlapping the motorcycle.

4.1.2 Validating the Training and Label Generation Pipeline

We validated the model architecture and the label generation pipeline in two sequential stages to systematically isolate potential sources of error.

First, to verify the integrity of the 3D detector and the optimisation hyperparameters, we trained the OV-SCAN Baseline from scratch using the OV-SCAN Reference Labels. By leveraging their offline data, we isolated the training mechanics from the label generation mechanics. As shown in Table 4.2, our training implementation, including the gradient accumulation adaptations, closely mirrors the published performance, achieving an overall mAP of 30.9 compared to the reported 31.1.

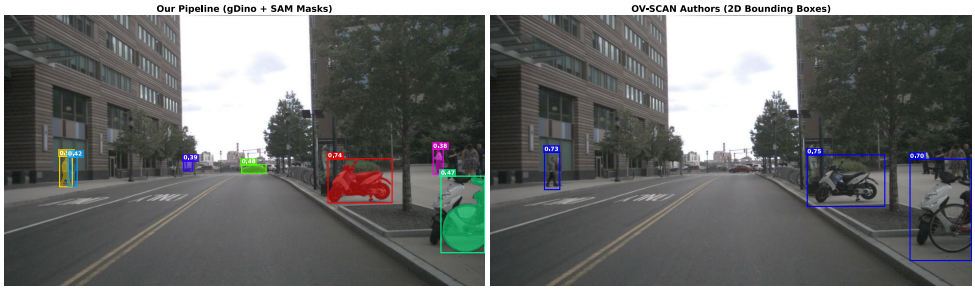


Figure 4.1: Validation of 2D visual proposals. Left: Our generated 2D bounding boxes and extracted SAM instance masks. Right: The original 2D bounding boxes from the OV-SCAN Reference Labels. While global confidence distributions vary due to undisclosed baseline parameters, salient objects (e.g., the motorcycle) demonstrate closely aligned scores (0.74 vs. 0.75), confirming the structural validity of the reproduction.

This negligible deviation of 0.2 mAP confirms that our core training loop is similar to the baseline. However, we observe a degradation in the Traffic Cone class (-3.0 mAP), likely attributable to non-deterministic training variance.

Second, to validate our complete reverse-engineered pipeline, we replaced the reference database with the Reproduced Global Labels generated entirely by our own reproduction of the Global Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) Pipeline. This full end-to-end reproduction achieves an overall mAP of 26.7.

While this 4.4 mAP difference relative to the published OV-SCAN results (31.1 mAP) confirms that offline data preparation affects downstream performance, this end-to-end comparison alone does not isolate which specific sub-component of the reconstructed pipeline drives the deviation. A class-wise analysis reveals that certain classes experience performance drops (e.g., `Car` decreases from 61.6 to 49.7), while other geometrically complex categories actually improve (e.g., `Bus` increases substantially from 22.0 to 32.4).

Crucially, this fully reproduced pipeline establishes a control baseline. By using this specific configuration as our reference point, we ensure a fair and direct comparison when evaluating the structural impact of the proposed Local Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Pipeline in the subsequent section.

Table 4.2: Validation of the training and label generation pipelines. All configurations use the OV-SCAN Baseline architecture, trained from scratch and evaluated under the Full Benchmark Protocol and compared against the published OV-SCAN results. C.V.: Construction Vehicle, T. C.: Traffic Cone.

Configuration	**all**	**Car**	**Truck**	**Bus**	**Trailer**	**C. V.**	**Pedestrian**	**Motorcycle**	**Bicycle**	**T. C.**	**Barrier**
OV-SCAN Baseline (Published)	31.1	61.6	30.3	22.0	0.6	6.8	60.1	39.8	31.0	39.6	18.8
OV-SCAN Reference Labels	30.9	61.6	30.0	22.7	0.8	5.0	59.7	40.2	32.1	36.6	20.7
Reproduced Global Labels	26.7	49.7	20.6	32.4	0.9	4.8	46.4	40.0	30.4	30.2	11.9

4.2 Evaluation of the Improved Pseudo-Label Generation

This section evaluates the proposed Local DBSCAN Pipeline. To isolate the impact of our architectural contribution, the proposed method is directly compared against the Global HDBSCAN Pipeline (Section 4.1.2), as both share the exact same underlying codebase and 2D vision pipeline.

4.2.1 Pseudo-Label Geometric Quality

A direct way to evaluate an offline pseudo-labelling pipeline prior to detector training is to measure the geometric accuracy of its generated bounding boxes before any neural network is trained. The proposals were evaluated against the full nuScenes ground truth using a 3D Intersection-over-Union (IoU) threshold of ≥ 0.5.

Table 4.3: Overall pseudo-label localisation quality against the nuScenes ground truth (3D IoU ≥ 0.5). Generated across the training set under the Full Benchmark Protocol, the local frustum constraint acts as a geometric filter, increasing both precision and recall.

Metric	Reproduced Global Labels	Local Labels
Overall Recall	11.1	**13.1**
Overall Precision	28.0	**34.9**

As shown in Table 4.3, transitioning from Reproduced Global Labels to the Local Labels improves geometric quality, yielding absolute gains of +2.0% in recall and +6.9% in precision.

To understand the mechanics driving these improvements, we must analyse the raw volume of generated labels per class. Figure 4.2 breaks down the proposal counts across both pipelines.

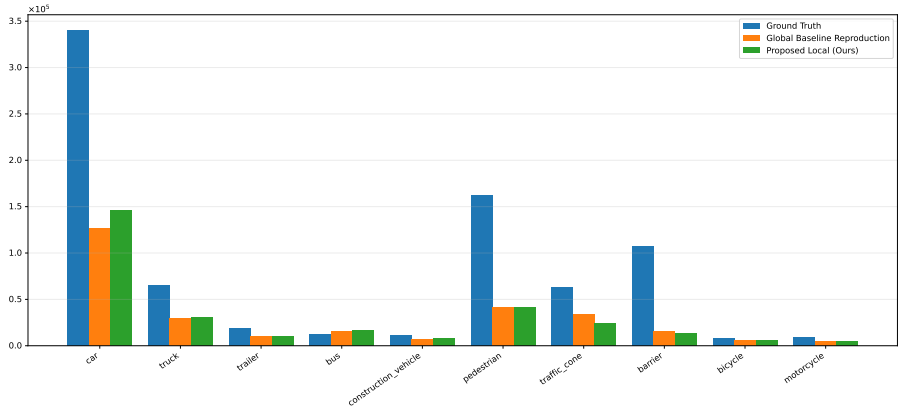


Figure 4.2: Class-wise label generation volumes. The local pipeline behaves distinctly across different object typologies, aggressively boosting large vehicle volumes while filtering out small objects.

The volume shifts reveal that the local frustum pipeline behaves distinctly depending on the object class. For large vehicles (e.g., `Car`, `Truck`, `Bus`), the local method recovers a significantly higher number of instances, yielding over 18,000 additional cars. Conversely, for small static street

clutter (`Traffic Cone`, `Barrier`), it generates far fewer proposals, producing nearly 9,500 fewer traffic cones compared to the global baseline. Finally, for Vulnerable Road Users (VRUs) (`Pedestrian`, `Bicycle`, `Motorcycle`), the local method yields a marginal decrease in total volume.

Crucially, these raw counts only highlight a difference in clustering behaviour. They do not inherently reflect the geometric correctness of the individual labels. To understand how these volume shifts translate into actual proposal quality, we must first visually inspect the spatial constraints of the clusters (Section 4.2.2) and subsequently evaluate their impact on downstream 3D detection (Section 4.2.4).

4.2.2 Qualitative Clustering Analysis

To visually contextualise these precision gains, we examine the underlying clustering behaviour on the raw LiDAR point clouds in Figure 4.3.

As shown in Figure 4.3b, the Global HDBSCAN Pipeline suffers from over-segmentation. It fractures the vehicle into disjointed sub-clusters, capturing only a small fraction of its true geometric mass. Consequently, the Particle Swarm Optimisation (PSO) fits a bounding box to this isolated fragment, resulting in an undersized and misaligned pseudo-label.

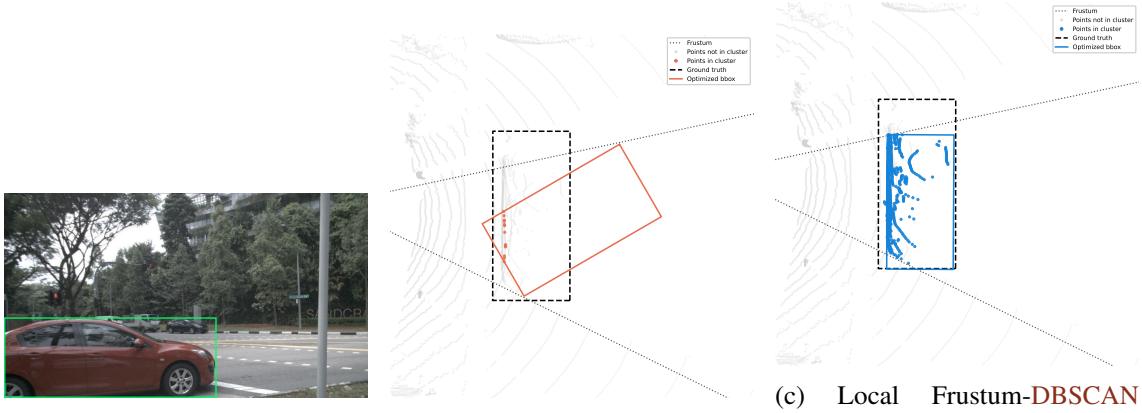
Conversely, the proposed local approach (Figure 4.3c) uses the 2D bounding box as a spatial boundary. This isolation allows local DBSCAN to safely "under-segment" the points within the frustum without the risk of absorbing adjacent background noise. By capturing a much denser, complete point cluster, the PSO converges to a tightly aligned bounding box that closely mirrors the ground truth.

While this local assumption is effective, it introduces a minor geometric limitation: because the 2D bounding box dictates the 3D search space, any physical occlusion or truncation in the camera view propagates to 3D. As seen in Figure 4.3c, the rear of the vehicle extends slightly outside the 2D image and is therefore truncated from the cluster. Nevertheless, the substantial overall improvements in precision suggest that relying on these 2D spatial priors is superior to the over-segmentation errors of global LiDAR clustering.

4.2.3 Computational Viability

To evaluate computational viability, a timing analysis was conducted on a subset of 2,727 frames. The initial 2D visual pipeline (Grounding DINO, SAM, and CLIP) dominates the latency, consuming an average of 4,798.5 ms per frame. The overall end-to-end execution time averages 8,362.0 ms for the Global HDBSCAN Pipeline and 8,454.1 ms for the Local DBSCAN Pipeline (with the minor differences between the sum of components and the total pipeline time attributed to I/O and format conversions).

Isolating the 3D-only processing times, which encompass ground removal, clustering, cross-modal association, and PSO, the Global HDBSCAN Pipeline averages 3,551.1 ms per frame, whereas the proposed Local DBSCAN Pipeline averages 3,643.2 ms, resulting, therefore, in a



(a) 2D Proposal (RGB) (b) Global HDBSCAN (Baseline) (c) Local Frustum-DBSCAN (Ours)

Figure 4.3: Comparative Bird's Eye View (BEV) analysis of a vehicle bounding box optimisation. **(a)** The initial 2D proposal. **(b)** The Global HDBSCAN Pipeline suffers from density-based over-segmentation, causing the PSO to optimise a poorly aligned bounding box. **(c)** Our proposed local frustum-DBSCAN safely clusters the majority of the vehicle's mass, leading to a highly accurate optimisation, despite a minor truncation at the rear edge caused by the 2D boundaries.

marginal overhead. Despite executing DBSCAN iteratively for every 2D frustum, the computational penalty remains negligible because the frustum acts as a spatial filter that reduces the number of points processed per clustering iteration.

4.2.4 Downstream 3D Detection Impact

To evaluate the downstream learning benefits of the improved proposal generator, both the OV-SCAN baseline and the proposed OV-D3D architecture were trained on the Reproduced Global Labels and our Local Labels under the Full Benchmark Protocol.

Table 4.4: 3D detection performance (mAP and NDS) trained on Reproduced Global Labels vs. Local Labels under the Full Benchmark Protocol. Both architectures exhibit performance improvements under localised pseudo-labelling.

Training Labels	**OV-SCAN (Baseline Architecture)**		**OV-D3D (Our Architecture)**	
	mAP ↑	**NDS** ↑	**mAP** ↑	**NDS** ↑
Global Labels (Baseline)	26.7	31.5	26.4	31.0
Local Labels (Ours)	**28.8** (+2.1)	**35.3** (+3.8)	**27.4** (+1.0)	**32.0** (+1.0)

As shown in Table 4.4, substituting the Reproduced Global Labels with the Local Labels yields consistent improvements across both architectures (+2.1 mAP / +3.8 NDS for OV-SCAN, and +1.0 mAP / +1.0 NDS for OV-D3D). While both models benefit from the cleaner supervisory signal, OV-SCAN exhibits a larger performance recovery. A plausible hypothesis compatible with these results is that multi-channel architectures leverage category-specific priors during bounding box regression, rendering them more responsive when supervisory noise is reduced, whereas single-channel proposals rely on geometric cues.

A class-wise breakdown (Table 4.5) further contextualises this behaviour against the generated label volumes (Section 4.2.1). For non-rigid VRUs (`Pedestrian`, `Bicycle`, `Motorcycle`), generating fewer but geometrically cleaner proposals yielded mAP increases across both architectures (up to +6.3 mAP on `Pedestrian` for OV-SCAN). Conversely, for small static objects such as `Traffic Cone`, where local frustum filtering reduced label volume by 28.3%, OV-SCAN improved (+6.9 mAP) through its dedicated semantic channel, whereas OV-D3D experienced a decrease (-3.3 mAP), suggesting that single-channel discovery is sensitive to volume reductions on small geometries lacking dedicated channels.

Table 4.5: Class-wise mAP shifts when training on Reproduced Global Labels vs. Local Labels under the Full Benchmark Protocol.

Architecture (Δ mAP)	**Car**	**Truck**	**Bus**	**Trailer**	**C. V.**	**Pedestrian**	**Motorcycle**	**Bicycle**	**T. C.**	**Barrier**
OV-SCAN (Local Labels - Global Labels)	-0.5	+1.7	+0.1	-0.1	+0.4	+6.3	+2.5	+3.7	+6.9	-0.9
OV-D3D (Local Labels - Global Labels)	-0.3	+0.7	+4.6	-0.3	-0.8	+3.6	+2.5	+3.9	-3.3	-1.1

In summary, training on the Proposed Local Labels yields consistent downstream gains across both network configurations, supported by direct comparisons against the Reproduced Global Labels control under identical training settings. However, because the localised pipeline transitions simultaneously from scene-wide clustering to camera frustums and from HDBSCAN to DBSCAN, these results demonstrate the aggregate benefit of the compound intervention rather than isolating the spatial bounding from the algorithm swap.

4.3 Our Improved Architecture: OV-D3D[1]

This section evaluates the effectiveness of the proposed OV-D3D architecture against the state-of-the-art OV-SCAN baseline on macro-level detection performance, robustness to sensor degradation, and, finally, zero-shot capabilities to isolate the respective bottlenecks of geometric extraction and semantic alignment.

4.3.1 Macro-Level 3D Detection Performance

To establish general detection capabilities, we first compare OV-D3D against existing 3D Open-Vocabulary Object Detection (OV-OD) methods under the Full Benchmark Protocol on the full nuScenes [6] validation set.

As shown in Table 4.6, OV-D3D achieves an overall mAP of 28.8 and an NDS of 31.7. While it outperforms earlier pseudo-labelling methods such as CoDA (10.3 mAP) and OpenSight (22.9 mAP), we observe an overall -2.3 mAP reduction compared to the published category-dependent OV-SCAN baseline (or -2.1 mAP relative to our 30.9 mAP global reproduction under identical reference labels).

[1] Portions of the text in this section have been adapted from the workshop paper: Angélico, P., et al., "Decoupling Geometry and Semantics for Robust Open-Vocabulary 3D Object Detection" [1], accepted at the European Conference on Computer Vision (ECCV) 2026.

Table 4.6: 3D Object Detection performance under the Full Benchmark Protocol. OV-D3D is supervised by the OV-SCAN Reference Labels. Removing category-specific heatmap channels eliminates semantic size shortcuts, leading to expected mAP trade-offs on some classes. However, our approach remains competitive across the vocabulary, matching the baseline on several primary classes and achieving performance gains on `Traffic Cone` and `Barrier`. *CoDA uses ground truth annotations for the `Car` class. C.V.: Construction Vehicle, T. C.: Traffic Cone.

Method	mAP ↑	NDS ↑	Car	Truck	Bus	Trailer	C. V.	Pedestrian	Motorcycle	Bicycle	T. C.	Barrier
OV-3DET [43]	5.7	12.0	16.1	2.9	1.3	0.0	0.3	21.1	2.0	1.8	11.0	0.3
CoDA [7]	10.3	16.3	85.5*	1.9	0.0	0.0	0.0	3.5	2.0	5.9	3.9	0.0
UP-VL [48]	11.3	17.2	19.4	7.9	3.1	0.4	4.3	30.8	16.8	14.1	15.6	0.5
Find n' Prop. [15]	16.7	22.4	24.3	8.6	11.1	0.1	4.1	22.8	35.8	34.7	21.3	4.4
OpenSight [79]	22.9	23.3	25.3	11.6	5.1	0.8	8.7	52.5	30.4	26.1	25.6	42.2
OV-SCAN [11]	**31.1**	**32.8**	**61.6**	**30.3**	**22.0**	0.6	**6.8**	**60.1**	39.8	**31.0**	39.6	18.8
OV-D3D (Ours)	28.8	31.7	60.3	20.5	14.9	**0.8**	5.2	53.3	**40.5**	27.2	**43.3**	**22.6**

This reduction suggests an expected architectural trade-off for achieving true vocabulary independence. Collapsing a 10-channel semantic heatmap into a single class-agnostic channel fundamentally alters the query initialisation phase. In the OV-SCAN baseline, queries are generated by summing geometric BEV features with one-hot encoded semantic embeddings, injecting rigid, category-specific size priors into the network before regression occurs.

By removing these explicit semantic embeddings, our 1-channel formulation deprives the regression head of these categorical shortcuts. Consequently, performance varies across object typologies, a pattern compatible with two structural hypotheses:

- **The Loss of Semantic Size Priors:** Classes characterised by high intra-class scale variance (e.g., *Truck*) or LiDAR sparsity (e.g., *Pedestrian*, *Bicycle*) experience Mean Average Precision (mAP) reductions. In a multi-channel baseline, predicting category-specific heatmaps directly from LiDAR BEV representations allows the network to associate spatial peaks with learnt and category-specific dimensional anchors. Consequently, once a proposal activates a specific class channel, the regression head refines the class-average bounding box geometry, providing a stable geometric prior even under degraded point returns. Deprived of these categorical routing cues, our class-agnostic formulation must regress 3D volumetric extents purely from geometric spatial features without dedicated scale priors, an inherently harder regression task that increases variance on scale-sensitive geometries.

- **Feature Consolidation on Ambiguous Geometries:** Conversely, the proposed formulation yields distinct improvements on small, rigid geometries such as `Traffic Cone` (+3.7 mAP) and `Barrier` (+3.8 mAP). In a multi-channel architecture, the weak spatial evidence for semantically ambiguous objects may bleed across multiple similar semantic channels (e.g., `Traffic Cone` vs. `Pedestrian` vs. `Barrier`). This channel competition fragments the probability mass, spawning multiple, conflicting predictions at the same spatial location. The 1-channel heatmap appears to mitigate this channel competition by consolidating all local spatial evidence into a single, cohesive objectness peak, generating cleaner initial queries for objects that do not require complex size priors to be accurately bounded.

4.3.2 Robustness to Spatial Degradation at Range

To evaluate generalisation under sensor degradation, we measure detection performance across cumulative distance intervals (Near, Mid, Far). Preliminary evaluations of this architecture [1] assessed robustness using visibility and occlusion stratifications. However, a subsequent audit identified an evaluation artefact in that protocol: while ground-truth annotations contain specific visibility labels, the network's spatial predictions do not. Filtering the ground truth by visibility without a corresponding way to filter the predictions skewed the resulting metrics. Consequently, this dissertation evaluates spatial degradation using cumulative distance from the sensor, an attribute that can be independently calculated for both predictions and ground-truth objects. Because LiDAR pulses diverge radially in autonomous driving environments, stratifying performance by distance serves as a consistent proxy for spatial degradation.

Table 4.7: Comparison of detection robustness across cumulative distance intervals (mAP %) evaluated under the Full Benchmark Protocol. OV-D3D is supervised by the OV-SCAN Reference Labels to isolate architectural behaviour from pseudo-label variance, and results are compared against the original OV-SCAN published values. While the category-dependent baseline achieves higher absolute mAP by leveraging semantic size priors, the class-agnostic OV-D3D architecture exhibits expected performance trade-offs under spatial degradation.

	Distance (mAP %)		
Method	**Near (0–18 m)**	**Mid (0–34 m)**	**Far (0–54 m)**
OV-SCAN Baseline	**48.1**	**35.6**	**27.7**
OV-D3D (Ours)	45.0	33.0	25.5

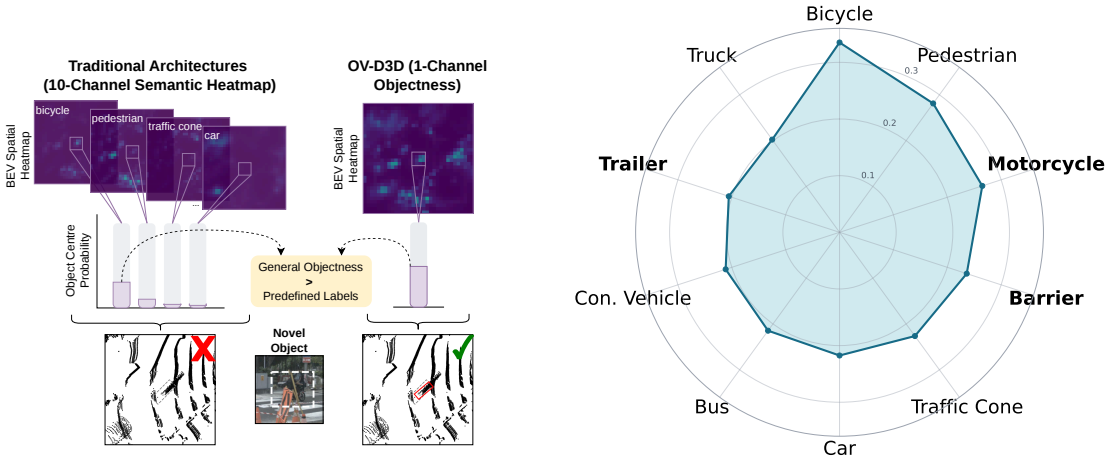
As detailed in Table 4.7, consistent with the overall results, OV-SCAN maintains a higher absolute mAP across all distance intervals. This is expected, as its architecture leverages closed-set semantic priors in the query initialisation phase to guide 3D bounding box regression.

However, an analysis of the relative performance decay reveals a critical architectural insight. When moving from the near to the far interval, OV-SCAN experiences a 42.4% relative mAP reduction, while OV-D3D exhibits a nearly identical 43.3% reduction. Because both architectures degrade at similar rates despite the presence or absence of semantic channels, these results indicate that intermediate semantic classification does not inherently resolve the challenge of spatial degradation at long ranges. Instead, the semantic prior acts as a relatively constant bounding-box regression aid across all distances.

4.3.3 The Zero-Shot Paradox

When evaluated on the novel holdout classes under the Zero-Shot Holdout Protocol (Section 3.6.3), both the OV-SCAN baseline and OV-D3D yield a score of 0 mAP. Because the nuScenes benchmark disregards precision and recall regions below 0.1, this outcome reflects a classification failure according to the official protocol, without necessarily precluding isolated true-positive detections.

To determine whether the architectures failed to extract the physical geometry or were instead bottlenecked by semantic alignment, we examine a concrete zero-shot instance: an unannotated `Motorcycle` (Figure 4.4).



(a) Object centre probability heatmaps. OV-SCAN splits confidence across predefined classes, while OV-D3D generates a single, more confident, peak.

(b) Zero-shot CLIP cosine similarities of the embedding of the discovered novel object by OV-D3D. *Novel classes are in bold.*

Figure 4.4: Analysis of a zero-shot `Motorcycle` instance. Although OV-D3D localises the novel object (a), the semantic alignment (b) is bottlenecked by visually similar base classes like `Bicycle` and `Pedestrian`.

As shown in Figure 4.4a, OV-SCAN's multi-channel architecture fractures the unannotated geometry. At the exact position of the novel object's centre, its probability mass is split across the "bicycle" (30.5%) and "pedestrian" (10.2%) channels, resulting in a weak, dispersed peak. In contrast, OV-D3D consolidates the exact same geometry into a single, confident objectness peak (42.1%). Yet, the raw cosine similarities (Figure 4.4b) reveal that the resulting embedding is masked by base classes that compose the object's visual appearance ("bicycle": 0.335, "pedestrian": 0.281).

This paradox suggests that the 0 mAP on novel classes stems largely from a semantic classification failure rather than a geometric one. We systematically isolate these two variables in the following sections.

4.3.4 Isolating Geometric Localisation

To evaluate geometric discovery independently of classification accuracy, we evaluate the final, regressed bounding boxes using Class-Agnostic Recall. To prevent the artificial inflation of recall metrics, all evaluations across all architectures are constrained by a maximum extraction of 200 region proposals per frame. This ensures that recall gains reflect a genuine increase in high-confidence object discovery, rather than a mere over-generation of low-confidence noise.

To evaluate macro-discovery, we isolate the 2.0 m centre-distance matching threshold. As detailed in Appendix D.1, OV-D3D achieves a higher global proposal recall on the standard training

protocol (78.35% vs. 74.72% for OV-SCAN). However, the architectural divergence becomes critical under the zero-shot holdout condition, where OV-D3D maintains 76.70% global proposal recall compared to 70.29% for the baseline.

Table 4.8: Class-Agnostic Recall (%) evaluated using a 2.0 m centre-distance threshold. Both models are supervised by the OV-SCAN Reference Labels. The table compares performance when the network is trained on the Full Benchmark Protocol (**Standard**) versus when target geometries are omitted from the training vocabulary (**Holdout**). The multi-channel baseline suffers a recall collapse, whereas OV-D3D maintains high geometric discovery.

	Barrier		**Motorcycle**		**Trailer**	
Method	Standard	Holdout	Standard	Holdout	Standard	Holdout
OV-SCAN Baseline	70.2	41.7	81.7	77.7	25.7	23.0
OV-D3D (Ours)	**77.8**	**65.6**	**88.2**	**87.6**	**34.1**	**30.8**

As detailed in Table 4.8, OV-D3D achieves higher geometric recall across all novel geometries. Most notably, when `Barrier` instances are removed from the training vocabulary, OV-SCAN's recall collapses from 70.2% (Standard) to 41.7% (Holdout). This performance delta exposes the limitation of multi-channel architectures: geometries that do not strongly activate a predefined base-class channel are treated as background noise and suppressed. By contrast, the OV-D3D architecture discovers these objects, yielding 65.6% recall on `Barrier`.

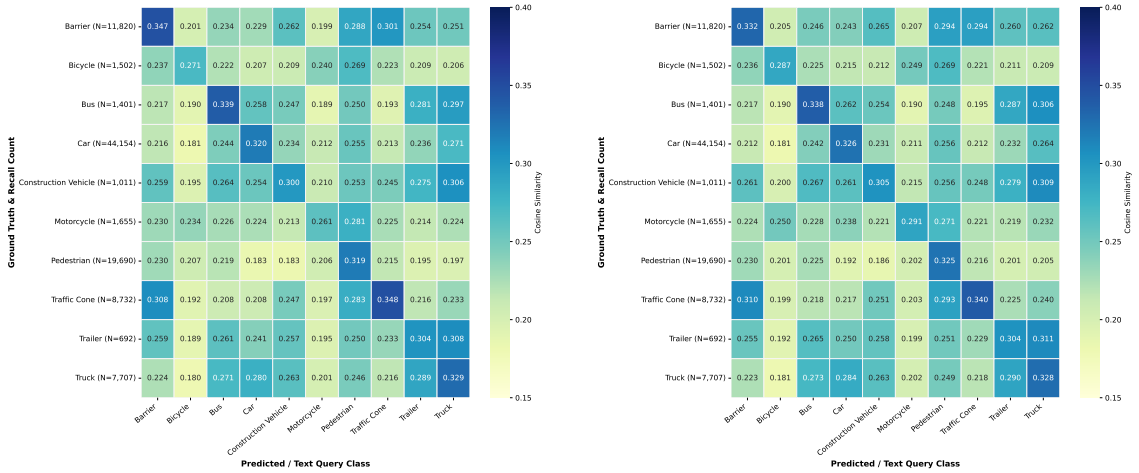
While we highlight the 2.0 m threshold as a standard metric for macro-discovery, analysing performance across stricter distance thresholds reveals the physical trade-offs of 1-channel architectures. As detailed in Appendix D.2, at highly strict matching thresholds (e.g., AR@0.5m), the multi-channel baseline yields higher recall. This indicates that semantic channels act as dimensional priors, helping traditional networks pinpoint a more accurate micro-centre of an object. Without these priors, OV-D3D trades micro-localisation precision for superior macro-discovery of novel shapes at extended thresholds (AR@2.0m and AR@4.0m).

In summary, under the Zero-Shot Holdout Protocol, single-channel formulations preserve proposal discovery across unannotated geometries, supported by higher holdout recall than the multi-channel baseline at standard macro-distance thresholds (e.g., 65.6% vs. 41.7% on `Barrier` and 87.6% vs. 77.7% on `Motorcycle` at 2.0 m). However, the extended analysis across thresholds (Appendix D.2) reveals that the 1-channel control slightly outperforms the full OV-D3D architecture on novel-class discovery (68.68% vs. 68.46% holdout mAR globally, and 49.91% vs. 48.78% on `Barrier`). This indicates that neither architecture represents an absolute winner across all objectives: the minimal 1-channel control preserves slightly higher agnostic discovery on novel shapes, whereas the full OV-D3D architecture better optimises global benchmark performance across both base and novel scenes.

4.3.5 The Semantic Alignment Bottleneck

Given that OV-D3D localises a substantial fraction of unannotated holdout instances at the 2.0 m threshold (up to 87.6% agnostic recall on `Motorcycle`), the 0 mAP score under the official

evaluation protocol points primarily to a bottleneck in cross-modal semantic classification rather than proposal generation.



(a) OV-SCAN Raw Cosine Similarities (b) OV-D3D Raw Cosine Similarities (Ours)

Figure 4.5: Complete semantic confusion matrices representing **Raw Cosine similarities** on the Full Benchmark Protocol (Columns = Text Queries, Rows = Ground Truth Objects). Notice the narrow numerical band across all methods, highlighting the fundamental ambiguity in the Vision-Language Model (VLM)'s feature space before logit scaling.

By analysing the raw semantic interactions in Figure 4.5, we can extract several critical insights regarding how current cross-modal architectures behave:

Feature Pollution via Co-occurrence (Column Analysis). Analysing the matrices by column reveals how specific text queries map to physical ground-truth objects. Ideally, a text query should exhibit high similarity with its target class. However, the data reveals significant semantic bleeding. For instance, the text query "barrier" exhibits a cosine similarity of 0.33 with actual `Barrier` objects, but it also maintains high similarities with `Traffic Cone` (0.31) and `Construction Vehicle` (0.26).

This indicates a limitation in the 2D VLM embeddings used for supervision. In autonomous driving datasets, barriers frequently co-occur (and physically overlap) with traffic cones and construction vehicles. Because the 2D VLM struggles to disentangle heavily occluded objects from their immediate surroundings, the resulting text-to-3D embedding space becomes polluted by contextual co-occurrence, forcing the network to confuse objects that share typical spatial environments.

The Asymmetry of Compositional Objects (Row Analysis). Analysing the matrices by row (how a specific physical object triggers various text queries) exposes an inherent asymmetry in cross-modal alignment.

Consider the `Motorcycle` class. When querying a ground-truth `Motorcycle` object, the similarity to the "pedestrian" text query is high (0.27). Conversely, when querying a ground-truth `Pedestrian` object, the similarity to the "motorcycle" text query drops (0.20).

This asymmetry reflects a structural bias in both the dataset and the VLM. Physically, a motorcycle in an outdoor dataset almost always includes a human rider. Therefore, its extracted 3D embedding is inherently compositional, bridging both the vehicle frame and the human. However, the majority of pedestrians are simply walking, unattached to vehicles. Consequently, the network cannot cleanly isolate the "motorcycle" concept from the "pedestrian" concept because the visual features of the former are entangled with the latter.

The Holdout Collapse and Architectural Divergence. When we evaluate the models under the Zero-Shot Holdout Protocol (where the true novel classes are excluded from the training vocabulary and pseudo-labels), the zero-shot mAP collapses to 0. During this evaluation, the models are provided with the full text vocabulary (including the novel queries) to test generalisation. Because class assignment is performed simply via the maximum cosine similarity (argmax), the underlying continuous similarities drive the following fallback behaviour:

- `Trailer`: Both OV-D3D and the OV-SCAN Baseline fall back to the "truck" query (0.298 and 0.297, respectively), due to their strong visual and geometric similarities.
- `Motorcycle`: Deprived of the "motorcycle" concept during training, OV-D3D falls back to the object's compositional sub-parts, registering highest similarities with "pedestrian" (0.285) and "bicycle" (0.254).
- `Barrier`: An architectural divergence occurs here. The OV-SCAN Baseline's highest fallback is "traffic cone" (0.291) over "pedestrian" (0.279). OV-D3D, however, falls back to "pedestrian" (0.288) over "traffic cone" (0.280). Because cross-modal alignment is supervised via 2D vision-language distillation, this divergence reflects susceptibility to visual co-occurrence in camera crops, where pedestrians and roadside barriers frequently coincide. Without category-specific routing channels to constrain feature attribution during training, single-channel proposals are slightly more prone to absorbing these entangled 2D contextual priors from the teacher model.

Semantic Alignment and Continuous Similarities. Across all three novel classes, the continuous latent representations preserve semantic proximity to logical visual fallbacks. However, the margins between the winning class and alternative candidates are narrow (for example, 0.285 for "pedestrian" versus 0.254 for "bicycle" on a novel `Motorcycle` instance).

Because the classification decision relies on the maximum cosine similarity (argmax), the detector assigns a single discrete label even when multiple concepts are closely matched. When visualising these scores through a temperature-scaled Softmax to interpret the relative distribution, the function amplifies these small margins into an apparently one-hot distribution, creating the visual

impression of an unambiguous classification. In reality, the underlying continuous embedding preserves partial semantic alignment with the object's constituent visual traits. The zero-shot classification collapse stems from the representation space itself, where contextual co-occurrence and visual overlap cause a base category to edge out the novel concept by a small numerical margin.

In exploratory trials, we investigated whether this semantic entanglement could be mitigated by enriching the training signal with broader linguistic priors. Specifically, we experimented with substituting the standard cosine embedding loss with a contrastive formulation against a wider vocabulary (LVIS [19]) to discourage semantic collapse onto dominant base classes. However, these preliminary runs showed no measurable improvement, consistently yielding 0 mAP on the novel holdout classes.

In summary, continuous cross-modal representations preserve partial semantic affinities to constituent visual sub-features, but fall short of correct zero-shot classification under standard benchmark metrics. This conclusion is supported by the distribution of unscaled cosine similarities, where novel classes consistently yield high affinities with visually co-occurring base categories. However, these observations do not demonstrate that vision-language distillation is inherently incapable of zero-shot transfer. Instead they show that unconstrained 2D distillation without region disentanglement transfers contextual visual co-occurrence directly into 3D features.

4.4 Ablation Studies[2]

To understand the impact of individual pipeline components, we conducted several ablation studies.

4.4.1 Impact of Prompt Engineering Strategies

The original baseline did not disclose the exact text prompts used to query the Grounding DINO [39] foundation model. To determine the most effective setup, we evaluated three vocabulary sets: the standard 10 nuScenes classes (*Default*), the fine-grained sub-classes from the baseline paper (*Specific*), and a reverse-engineered vocabulary from the authors' reference database (*Authors*). Each was tested using two prompting strategies: iterating through the list with N individual prompts versus evaluating a single concatenated string. For both strategies, a global Non-Maximum Suppression (NMS) is then applied.

The quantitative trade-offs of these prompt strategies are visually corroborated in Figure 4.6. While Table 4.9 indicates that the *Default + Single Prompt* strategy generates fewer proposals on average across the entire dataset, Figure 4.6 (Left) illustrates the high precision of this extraction: it reliably localises distant objects while remaining robust to semantic noise.

Conversely, forcing the VLM to evaluate highly granular sub-classes via the *Specific + N Prompts* strategy (Right) actively induces semantic hallucination, yielding confident false positives for irrelevant categories such as a "portable message board". Although evaluating N independent

[2]Portions of the text in this section have been adapted from the workshop paper: Angélico, P., et al., "Decoupling Geometry and Semantics for Robust Open-Vocabulary 3D Object Detection" [1], accepted at the European Conference on Computer Vision (ECCV) 2026.

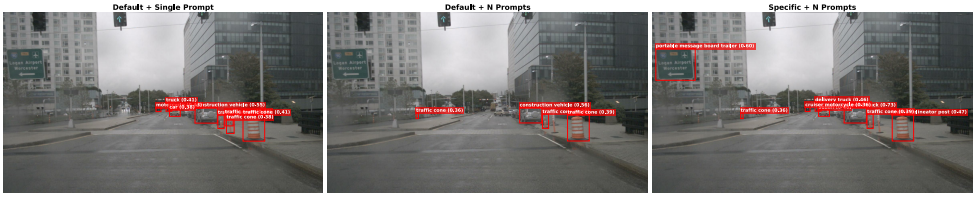


Figure 4.6: Qualitative impact of 2D prompt engineering strategies using Grounding DINO. Left: Default classes (Single Prompt). Middle: Default classes (*N* Prompts). Right: Specific fine-grained classes (*N* Prompts). While independent prompting (Middle) successfully recovers challenging small objects like the traffic cone, the highly granular *Specific* vocabulary (Right) induces semantic hallucination, generating a false positive for an irrelevant sub-class ("portable message board"). The *Default + Single Prompt* approach (Left) maintains high precision for distant objects while inherently suppressing semantic false positives.

prompts (Middle) marginally improves spatial recall for challenging, small-scale items, successfully detecting a traffic cone missed by the single prompt, this localised gain does not justify the associated computational bottleneck nor the global increase in false positive rates. Consequently, the *Default + Single Prompt* configuration was adopted as a practical trade-off between proposal precision and computational runtime.

Table 4.9: Comparison of 2D prompt engineering strategies for Grounding DINO. Metrics are percentages. *N prompts* executes a separate forward pass per class, while *Single Prompt* uses a concatenated vocabulary string.

		Recall @ IoU			Score MAE @ IoU			Avg Proposals/Img
Classes	Strategy	0.5	0.75	0.9	0.5	0.75	0.9	
Default	Single	31.99	15.82	2.16	46.31	46.20	46.86	2.86
Default	N prompts	34.56	16.49	2.09	39.19	37.40	35.58	4.11
Specific	Single	24.18	13.96	1.91	57.80	57.81	57.69	2.47
Specific	N prompts	**36.08**	**17.36**	**2.25**	**34.41**	**31.35**	**28.61**	6.12
Authors	Single	11.78	5.04	0.60	50.59	50.11	50.47	2.94
Authors	N prompts	26.85	12.68	1.67	44.59	44.58	43.59	5.35

As detailed in Table 4.9, employing *N* individual prompts consistently yielded higher bounding box recall across all vocabulary sets. However, this strategy introduces a computational bottleneck, requiring *N* independent forward passes through the Grounding DINO [39] architecture per image.

Conversely, the *Default + Single Prompt* strategy requires only one forward pass per image, vastly accelerating the pipeline. While exhibiting a slight drop in absolute recall ($\sim 4\%$ at IoU 0.50), it maintains highly efficient extraction (2.86 proposals per image). Moreover, employing a concatenated string inherently resolves semantic redundancy among different proposals. Consequently, this strategy was selected for the pipeline reproduction due to its simplicity, efficiency and lack of further details on the original paper.

4.4.2 Sensitivity of the 3D Pipeline

The 3D proposal pipeline relies on a sequence of heuristic thresholds (e.g., HDBSCAN density, spatial matching distances, and NMS overlap). To evaluate the sensitivity of these heuristics and select an operational configuration, we conducted a hyperparameter grid search (the comprehensive results are provided in Appendix B.3).

Table 4.10: Sensitivity summary of the 3D pipeline hyperparameters, evaluated under the Representative Subset Protocol. The degradation in F1 score underscores the pipeline's sensitivity to heuristic clustering thresholds.

Grid Search State	**Ours vs. Ground Truth**			**Ours vs. OV-SCAN Ref. Labels**		
	Recall	Precision	F1	Recall	Precision	F1
Upper Bound (Max Observed)	13.3	29.0	17.8	16.8	28.5	20.7
Selected Configuration	13.0	27.8	17.8	16.0	28.5	20.5
Lower Bound (Min Observed)	8.2	21.9	12.4	12.8	18.7	15.5

As summarised in Table 4.10, the grid search reveals that the global 3D pipeline is sensitive to miscalibration, with the F1-Score dropping (from 17.8% to 12.4%) under suboptimal configurations. While our Selected Configuration isolates a stable performance plateau that approximates the reference labels, the inability to push recall higher indicates that the upper bound is constrained by undocumented architectural nuances in the published paper.

Crucially, it must be noted that this grid search was conducted under the Representative Subset Protocol. As previously presented in Table 4.3, when these hyperparameters are applied to the Full Benchmark Protocol, the global pipeline's recall degrades from 13.0% to 11.1%. This highlights the inherent fragility of spatial heuristics.

To contextualise the precision variance observed in Table 4.10, Figure 4.7 provides a BEV comparison of the generated 3D proposals. Visual inspection confirms our pipeline exhibits a higher false positive rate. This over-generation of unaligned proposals directly accounts for the precision degradation, emphasising the sensitivity of the baseline's parameters and clustering algorithms.

Optimisation Convergence. Beyond spatial heuristics, the computational bottleneck of the 3D pipeline lies in the PSO bounding box fitting. The original authors used 3000 iterations per proposal. To evaluate the convergence of this module, we ablated the iteration count. Performance is measured using bounding box precision and recall at a 3D IoU threshold of ≥ 0.50.

As demonstrated in Table 4.11, our intermediate metrics indicate no meaningful performance degradation when drastically reducing the iteration count. In fact, executing the swarm for 500 iterations yields marginally higher F1-scores than the 3000 iterations utilised by the authors. This accelerated convergence behaviour is likely attributed to our adherence to the theoretical optimisation constraints defined by Engelbrecht [14] as detailed in Section 3.2. Because the authors' exact formulation remains undisclosed, these results indicate that our implementation achieves comparable box quality at a fraction of the baseline's computational cost.

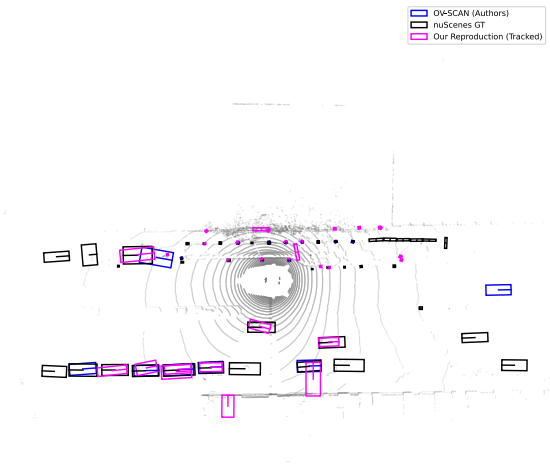


Figure 4.7: Qualitative comparison of 3D region proposals in BEV. Ground Truth bounding boxes are compared against the OV-SCAN Reference Labels (Authors) and Reproduced Global Labels (Ours). While our pipeline captures the spatial localisation of true positive objects, it generates a higher density of unaligned proposals (false positives) in cluttered regions.

Table 4.11: Impact of Particle Swarm Optimisation (PSO) iteration count on final 3D bounding box quality (evaluated at IoU ≥ 0.50) under the Representative Subset Protocol.

	Ours vs. Ground Truth			Ours vs. OV-SCAN Ref. Labels		
Iterations	Recall	Precision	F1	Recall	Precision	F1
OV-SCAN Ref. Labels vs GT	13.3	50.4	21.0	-	-	-
3000	12.8	27.5	17.5	15.4	27.3	19.7
1500	12.9	27.6	17.6	15.5	27.6	19.8
500	**13.0**	**27.8**	**17.8**	**16.0**	**28.5**	**20.5**

Ultimately, we selected a computationally efficient configuration ($\varepsilon = 0.8$, $N_{\text{matched}} = 1$, NMS IoU=0.50, and 500 PSO iterations) that achieved parity with the upper bounds while minimising processing overhead.

The Effect of Temporal Tracking. We evaluated the impact of integrating ImmortalTracker [66] to enforce temporal consistency, testing three configurations: *None* (tracker disabled), *Match + Delete Noisy* (filtering transient proposals), and *Match + Delete Noisy + Interpolate* (injecting hallucinated boxes for occluded frames).

Table 4.12: Impact of ImmortalTracker configurations on final 3D detection mAP. Evaluated using the OV-SCAN Baseline architecture under the Representative Subset Protocol.

Tracker Functionality	**mAP (%)**
None (Tracker Disabled)	21.0
Match + Delete Noisy	**21.3**
Match + Delete Noisy + Interpolate	21.2

As demonstrated in Table 4.12, the integration of the temporal tracker yields negligible improvements (≤ 0.3 mAP). This lack of variance indicates that while the tracker successfully filters some transient noise, injecting interpolated bounding boxes into frames where no physical LiDAR points exist provides no meaningful geometric supervision for the 3D detector. Nevertheless, on the final reproduction, we used the tracker for matching and deleting boxes.

4.4.3 OV-D3D Component Ablations

To systematically evaluate the contribution of our proposed modules, we conducted an orthogonal component ablation on the full nuScenes dataset.

As a point of reference, the original OV-SCAN authors briefly reported achieving 25.2 mAP when naively collapsing their architecture to a 1-channel heatmap, though the specific parameters of that configuration remain undocumented. To establish a foundation for our experiments, we built a controlled 1-Channel Control architecture using a standard 3×3 convolutional heatmap, no explicit scale priors, and a simple Feed Forward Network (FFN) alignment head. This clean, minimalist configuration achieves 27.8 mAP, improving upon the performance trends observed in the single-channel ablation reported by the original authors, though, possibly, under differing experimental configurations.

From this stable foundation, we independently ablated our three architectural axes, focusing not only on overall performance but specifically on Mean Average Translation Error (mATE) and Mean Average Scale Error (mASE) to diagnose geometric bottlenecks. The results are summarised in Table 4.13.

The 1-Channel Control. The most significant insight from Table 4.13 lies in the performance of the 1-Channel Control. By simply stripping away the 10 categorical semantic channels and the

Table 4.13: Architectural Ablations. All configurations are supervised by the OV-SCAN Reference Labels and evaluated under the Full Benchmark Protocol. Starting from a 1-channel control model (27.8 mAP), we independently ablate dimension priors, spatial receptive fields, and alignment head complexity. We report mATE (centre translation error) and mASE (bounding box scale error) to isolate the structural impact of each module.

Ablation Axis	**Configuration**	**mAP ↑**	**NDS ↑**	**mATE ↓**	**mASE ↓**
-	OV-SCAN Baseline (Reported 1-Ch)	25.2	-	-	-
1-Channel Control	1-Ch + No Scale Prior + 3×3 + FFN	27.8	31.0	0.4903	0.3512
1. Scale Prior	+ Simple Scale	28.1	31.2	0.4936	0.3497
(w/ 3×3 + FFN)	+ Residual Scale	27.7	31.0	0.4912	0.3460
	+ Harmonic Scale	28.3	31.4	0.4915	0.3487
2. Receptive Field	+ Vanilla ASPP	28.2	31.3	0.4948	0.3513
(w/ Harmonic + FFN)	**+ Gated ASPP (Ours)**	**28.8**	**31.7**	**0.4752**	0.3494
3. Alignment Head	+ Local Self-Attention (SA)	28.6	31.6	0.4823	0.3479
(w/ Harmonic + 3×3)	+ Global SA	28.6	31.4	0.4983	0.3493

complex cross-attention alignment head, the 1-channel control architecture still achieves 27.8 mAP. This supports the premise that class-agnostic objectness provides a viable spatial foundation for 3D proposal generation without requiring intermediate categorical routing.

1. The Scale Prior Bottleneck. By removing the multi-class heatmap, the Transformer Decoder is deprived of the categorical size hints. We hypothesised that explicitly predicting these dimensions via a Dense Dimension Head and encoding them via Harmonic Scale Embeddings would mitigate this loss of bias. While adding Harmonic Embeddings provides a +0.5 mAP gain over the 1-Channel Control, a critical analysis of the error metrics reveals a stubborn bottleneck. The scale error (mASE) barely shifts, reducing only marginally from 0.3512 to 0.3487. This plateau highlights an inherent limitation in class-agnostic architectures: predicting the precise 3D volumetric extents of an object from a single, unclassified BEV pixel is an inherently ill-posed problem. Without knowing what the object is, the network struggles to confidently regress its boundaries, confirming that scale variance remains a challenge for open-vocabulary pipelines.

2. Dynamic Receptive Fields Fix Translation, Not Scale. To address the extreme scale variance hitting the 1-channel centre proposal mechanism, we replaced the standard 3×3 convolution with a Gated ASPP module. As shown in the data, this addition does not resolve the scale error (mASE remains stagnant at 0.3494). However, it improves centre localisation, driving the translation error (mATE) down from 0.4915 to 0.4752. This metric divergence illustrates the module's physical behaviour: dynamically adapting receptive fields successfully allows the network to pinpoint the spatial centres of vastly differently sized objects (e.g., pedestrians vs. buses) within the same channel, but it does not assist the regression head in estimating their final volumetric extents.

3. Alignment Head Complexity and Trade-offs. Finally, we ablated the projection mechanism responsible for cross-modal distillation. While replacing the simple FFN with Local SA yields

marginal improvements in centre localisation (reducing mATE to 0.4823), the overall detection gain is negligible given the added computational overhead. The same occurs with Global Self-Attention which marginally degrades localisation accuracy, causing the translation error (mATE) to grow to 0.4983.

By allowing queries to exchange unconstrained inter-object context globally, the attention mechanism likely blurs the spatial boundaries established by the geometric backbone. Because complex attention heads offer no definitive performance breakthroughs, we retain the simple FFN alignment head in our final architecture. This supports the engineering intuition that once 3D queries are geometrically stabilised by the backbone, projecting them into the VLM latent space should be treated as a translation task, rather than an over-parameterised spatial mixing operation.

Architectural Justification. While the 1-Channel Control provides a robust foundation for class-agnostic discovery, its lack of dimensional priors inherently limits bounding box regression accuracy (27.8 mAP). The full OV-D3D architecture introduces Harmonic Scale Embeddings and dynamic receptive fields to mitigate this bottleneck, providing soft dimensional hints without explicitly coupling the geometry to semantic classes.

To verify that these modules provide systemic geometric stabilisation rather than merely overfitting to dense, nearby objects, we evaluate the architectural difference across cumulative distance intervals.

Table 4.14: Comparison of detection robustness across cumulative distance intervals (mAP %) evaluated under the Full Benchmark Protocol. Both models are supervised by the OV-SCAN Reference Labels to isolate architectural behaviour from pseudo-label variance. We ablate the 1-Channel Control with the full OV-D3D architecture. The proposed modules yield consistent geometric stabilisation across all spatial ranges.

	Distance (mAP %)		
Method	**Near (0–18 m)**	**Mid (0–34 m)**	**Far (0–54 m)**
1-Channel Control	43.6	31.8	24.7
OV-D3D (Ours)	**45.0**	**33.0**	**25.5**

As detailed in Table 4.14, the integration of the proposed modules yields a modest but consistent performance uplift across the entire spatial range. OV-D3D achieves gains of $+1.4$ mAP in the Near interval, $+1.2$ mAP in the Mid interval, and $+0.8$ mAP in the highly degraded Far interval.

In summary, starting from a 1-channel control (27.8 mAP), integrating Harmonic Scale Embeddings and Gated ASPP recovers +1.0 mAP on the Full Benchmark Protocol, reaching 28.8 mAP under the OV-SCAN Reference Labels condition. This recovery is supported by isolated ablation axes tracking translation (mATE) and scale (mASE) errors. However, the observed mechanics depart from their original motivations: Gated Atrous Spatial Pyramid Pooling (ASPP) improves spatial centre localisation rather than volumetric scale, while Harmonic Scale Embeddings enhance general query quality with minimal effect on scale error. Furthermore, while OV-D3D achieves superior overall performance on base categories, the minimal 1-channel control retains a marginal

recall advantage on unannotated holdout classes. Neither configuration represents an absolute winner: the minimal control preserves raw discovery on unseen geometries, whereas the full OV-D3D configuration better balances overall detection across familiar driving scenes.

Chapter 5

Conclusion

This thesis investigated the core bottlenecks of 3D Open-Vocabulary Object Detection (OV-OD) in outdoor autonomous driving environments. To establish a baseline control, we integrated the architecture of OV-SCAN [11] and reverse-engineered its closed-source pseudo-labelling pipeline from literature descriptions to instantiate our Reproduced Global Labels. Building upon this infrastructure, we developed the Local Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Pipeline, which restricts spatial clustering to 2D camera frustums and replaces global hierarchical clustering with localised DBSCAN, yielding improvements in both pseudo-label precision and downstream detection Mean Average Precision (mAP). Finally, we investigated the architectural conditioning of query initialisation by proposing OV-D3D, an architecture that removes category-specific heatmaps in favour of a single objectness channel augmented with continuous scale and adaptive receptive-field modules. Our evaluations demonstrated that while category-independent proposal generation avoids closed-set routing, the removal of categorical anchors degrades bounding box regression precision on scale-sensitive geometries. Under the Zero-Shot Holdout Protocol, we showed that although single-channel objectness reliably discovers unannotated geometries, zero-shot performance remains fundamentally bottlenecked by cross-modal semantic alignment.

5.1 Answers to the Research Questions

RQ 1: How can information from 2D foundation models be used to constrain the heuristic errors inherent to LiDAR point cloud clustering, thereby improving the generation of reliable 3D pseudo-labels?

Main Evidence and Supported Finding. Information from 2D foundation models is leveraged by projecting 2D bounding boxes from Grounding DINO [39] into discrete 3D camera frustums and restricting clustering exclusively to the enclosed points. This restriction reduces the spatial heterogeneity handled by each clustering operation, enabling the replacement of the baseline's Global Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

with a localised DBSCAN implementation. Empirically, this localised bounding mitigates density-based over-segmentation on large objects. When evaluated against nuScenes ground truth at 3D IoU ≥ 0.5, the proposed pipeline improved pseudo-label precision by $+6.9\%$ (from 28.0% to 34.9%) and recall by $+2.0\%$ (from 11.1% to 13.1%). When used to supervise downstream 3D detectors under the Full Benchmark Protocol, these cleaner pseudo-labels yielded consistent performance improvements across architectures, achieving $+2.1$ mAP and $+3.8$ NDS on the OV-SCAN baseline, and $+1.0$ mAP and $+1.0$ NDS on OV-D3D.

Remaining Delimitation. Because the proposed pipeline simultaneously alters both the spatial clustering region (from the global scene to individual frustums) and the clustering algorithm (from HDBSCAN to DBSCAN), the empirical gains demonstrate the aggregate benefit of this compound intervention. The current experimental setup does not isolate the independent contribution of the spatial frustum constraint from that of the DBSCAN parameterisation. Furthermore, objects that intersect the image boundaries produce truncated camera frustums, which restrict bounding box fitting to partial geometries. Therefore, developing multi-camera frustum merging remains a relevant control for future work.

RQ 2: What is the impact on the detection performance of base classes when eliminating the network's reliance on intermediate closed-set classification?

Main Evidence and Supported Finding. Eliminating intermediate closed-set classification trades off overall bounding box regression precision on familiar base classes in exchange for vocabulary independence and the suppression of channel competition. On the full benchmark protocol supervised by reference labels, our minimal 1-Channel Control achieved 27.8 mAP, while the complete OV-D3D architecture reached 28.8 mAP (compared to 31.1 mAP for the category-dependent OV-SCAN baseline). The observed -2.3 mAP delta is driven primarily by the removal of categorical dimensional shortcuts: classes with high intra-class scale variance or LiDAR sparsity suffer from the absence of category-specific bounding-box anchors, leading to lower regression accuracy on categories such as `Truck` (-9.8 mAP), Buses (-7.1 mAP), and `Pedestrian` (-6.8 mAP). Conversely, eliminating category-specific heatmaps prevents weak spatial evidence from bleeding across competing semantic channels, consolidating objectness peaks to secure absolute gains on visually ambiguous geometries such as `Traffic Cone` ($+3.7$ mAP) and `Barrier` ($+3.8$ mAP).

Remaining Delimitation. While the core hypothesis tested the viability of single-channel proposal generation, the auxiliary modules introduced to recover lost structural cues exhibited mechanics diverging from their original design motivations. Specifically, Gated Atrous Spatial Pyramid Pooling (ASPP) predominantly improved spatial centre localisation (reducing translation error Mean Average Translation Error (mATE) from 0.4915 to 0.4752) rather than resolving scale variance, while Harmonic Scale Embeddings improved general query feature quality ($+0.5$ mAP) with minimal effect on volumetric scale error (Mean Average Scale Error (mASE) remained at 0.3487).

Regressing precise 3D dimensions purely from geometric spatial features without categorical anchors remains an ill-posed challenge, and our findings reflect performance under a specific query-based architecture that may vary under alternate 3D detection backbones.

RQ 3: To what extent can 3D OV-OD architectures generalise to novel object geometries in real-world driving scenarios?

Main Evidence and Supported Finding. Under the Zero-Shot Holdout Protocol, where supervisory labels for novel categories are withheld during training while physical sensor returns remain in the background, evaluations revealed an empirical divergence between spatial discovery and semantic transfer. Due to the official nuScenes evaluation protocol discarding precision-recall regions below 0.1, both the baseline and OV-D3D scored 0 mAP on novel classes, reflecting a classification failure without necessarily precluding isolated true detections. Isolating spatial proposal generation via Class-Agnostic Recall demonstrates that single-channel objectness recovers unseen shapes across standard distance thresholds. At the primary 2.0 m centre-distance threshold, OV-D3D substantially outperforms the multi-channel baseline across all holdout classes, securing 65.61% vs. 41.71% on unannotated `Barrier`, 87.61% vs. 77.66% on `Motorcycle`, and 30.77% vs. 23.01% on `Trailer` (with a global dataset recall of 76.70% vs. 70.29%). Across distance thresholds ($\mathbb{D} = \{0.5, 1.0, 2.0, 4.0\}$ m), OV-D3D similarly retains higher overall holdout mean Agnostic Recall (68.46% vs. 65.31% mAR globally, 48.78% vs. 34.64% on `Barrier`, and 77.58% vs. 72.34% on `Motorcycle`). The multi-channel baseline collapses on these unannotated classes because geometries that do not activate predefined channels are suppressed as background. Diagnostic analysis of unscaled cosine similarities indicates that the downstream classification failure is semantic: continuous latent representations preserve proximity to visual sub-features (e.g., an unannotated `Motorcycle` registering highest affinity with its rider's pedestrian features at 0.285 similarity), but unconstrained 2D distillation allows contextual co-occurrence to dominate novel concepts.

Remaining Delimitation. Neither architecture represents an absolute winner across all zero-shot evaluation criteria. At the strictest micro-centre matching threshold (0.5 m), the multi-channel baseline maintains higher recall due to its rigid dimensional anchors, and the minimal 1-Channel Control slightly outperforms the full OV-D3D configuration in pure novel discovery (68.68% vs. 68.46% holdout mAR) by remaining entirely free of learnt scale priors. OV-D3D, by contrast, better balances global precision across the full benchmark (28.8 vs. 27.8 mAP). Furthermore, our experiments demonstrate that semantic misalignment occurs, but cannot isolate whether this failure originates during 3D latent distillation or is directly inherited from visual ambiguity in the upstream 2D Vision-Language Model (VLM) teacher.

5.2 Limitations and Future Work

Impact of Non-Deterministic Training Variance. Outdoor 3D OV-OD training pipelines incur substantial computational overhead, limiting the feasibility of extensive multi-seed training runs on available hardware. Future benchmark implementations should incorporate multi-seed statistical averaging to quantify non-deterministic variance and confirm fine-grained metric margins.

Anchor-Free Spatial Clustering. The offline pseudo-labelling pipeline relies on category-specific dimensional anchors to bound the continuous Particle Swarm Optimisation (PSO) bounding-box search. This reliance restricts the pipeline's capability to generalise to atypical or non-standard vehicle shapes. As documented in Appendix E, preliminary attempts to remove these dimensional bounds caused optimisation instability on sparse point clouds. Developing density-adaptive and anchor-free techniques represents an open challenge.

Temporal Aggregation of Pseudo-Labels. Despite constraining the 3D search space via 2D frustums, single-frame pseudo-labelling remains highly susceptible to momentary occlusion. Integrating 2D knowledge across time via temporal tracklets could further stabilise the 3D optimisation process. As detailed in Appendix E, our preliminary explorations into temporal aggregation showed potential for stabilising object dimensions, but revealed that current offline optimisation is highly sensitive to 2D heuristic tracking noise. Developing highly robust, 3D-aware association models to safely enable global tracklet optimisation is a critical next step.

Statistical Scope of Zero-Shot Evaluation. While the Zero-Shot Holdout Protocol exposed the divergence between geometric discovery and semantic transfer, the empirical scope was constrained to a single 3-class holdout split (`Motorcycle`, `Barrier`, `Trailer`) on the nuScenes dataset. Validating the generalisability of category-independent objectness across wider vocabulary partitions, diverse sensor configurations, and alternative driving benchmarks is necessary to map the domain limits of 3D OV-OD.

Isolating the 2D Teacher from 3D Distillation. A critical open question highlighted by this work is whether zero-shot semantic collapse stems from the 3D detector's inability to extract discriminative features from holistic bounding boxes, or whether it reflects ambiguities inherited from the pretrained 2D VLM teacher. Future research must establish isolated 2D zero-shot baselines on the exact camera crops to evaluate teacher fidelity directly. If visual overlap (such as riders on motorcycles) degrades 2D alignment, future architectures should explore part-level geometric distillation or contrastive loss formulations to prevent contextual visual co-occurrence from overriding novel concepts.

Continuous Representations for Perception. Prevailing open-vocabulary evaluation protocols require mapping continuous latent features to discrete text labels via hard argmax selection, discarding the rich affinity distributions encoded in continuous embedding spaces. Rather than forcing

early categorisation through discrete text matching, future research could explore interfacing continuous perception embeddings directly with downstream trajectory prediction and planning modules. Investigating how continuous multimodal representations can inform obstacle spatial extent and motion forecasting remains an important frontier for autonomous systems.

In summary, this work demonstrates that 3D geometric proposal discovery can function independently of predefined categorical routing, but achieving robust zero-shot open-vocabulary detection remains bounded by cross-modal semantic alignment. Addressing this alignment bottleneck and exploring continuous representations represent essential steps towards robust perception in autonomous systems.

References

[1] Pedro Angélico, José Monteiro, and Luis Teixeira. "Decoupling Geometry and Semantics for Robust Open-Vocabulary 3D Object Detection". In: *Eur. Conf. Comput. Vis. Worksh.* Springer Nature Switzerland, 2026.

[2] Claudine Badue et al. "Self-Driving Cars: A Survey". In: *Expert Systems with Applications* 165 (2021). DOI: 10.1016/j.eswa.2020.113816.

[3] Xuyang Bai et al. "TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2022, pp. 1080–1089. DOI: 10.1109/CVPR52688.2022.00116.

[4] Alex Bewley et al. "Simple online and realtime tracking". In: *IEEE Int. Conf. Image Process.* 2016, pp. 3464–3468. DOI: 10.1109/ICIP.2016.7533003.

[5] Corentin Bunel et al. "Addressing Open-set Object Detection for Autonomous Driving perception: A focus on road objects". In: *IEEE Int. Vehicles Symposium*. IEEE. 2024. DOI: 10.1109/IV55156.2024.10588525.

[6] Holger Caesar et al. "nuScenes: A Multimodal Dataset for Autonomous Driving". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020. DOI: 10.1109/CVPR42600.2020.01164.

[7] Yang Cao et al. "CoDA: collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection". In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. Curran Associates Inc., 2023. DOI: 10.52202/075280-3145.

[8] Mathilde Caron et al. "Emerging Properties in Self-Supervised Vision Transformers". In: *Int. Conf. Comput. Vis.* 2021. DOI: 10.1109/ICCV48922.2021.00951.

[9] Xiaozhi Chen et al. "Multi-view 3D Object Detection Network for Autonomous Driving". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 6526–6534. DOI: 10.1109/CVPR.2017.691.

[10] Xinlong Cheng and Lei Li. "Open 3D World in Autonomous Driving". In: *arXiv preprint arXiv:2408.10880* (2024). URL: https://arxiv.org/abs/2408.10880.

[11] Adrian Chow et al. "OV-SCAN: Semantically Consistent Alignment for Novel Object Discovery in Open-Vocabulary 3D Object Detection". In: *Int. Conf. Comput. Vis.* 2025, pp. 7990–8000. DOI: 10.1109/ICCV51701.2025.00749.

[12] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2016, pp. 3213–3223. DOI: 10.1109/CVPR.2016.350.

[13] Akshay Dhamija et al. "The Overlooked Elephant of Object Detection: Open Set". In: *IEEE Winter Conf. on App. of Comput. Vis.* 2020.

[14] Andries P. Engelbrecht. *Computational Intelligence: An Introduction*. 2nd. John Wiley, 2007. ISBN: 978-0-470-03561-0. DOI: 10.1002/9780470512517.

[15] Djamahl Etchegaray et al. “Find n’ Propagate: Open Vocabulary 3D Object Detection in Urban Scenes”. In: *Eur. Conf. Comput. Vis.* Springer, 2024. DOI: 10.1007/978-3-031-73661-2_8.

[16] Zhenyu Fang et al. “Cas-OVD: Cascaded Open-Vocabulary Detection of Small Objects Using Multi-Refined Region Proposal Network in Autonomous Driving”. In: *IEEE Trans. Multimedia*. Vol. 28. 2026, pp. 757–771. DOI: 10.1109/TMM.2025.3632649.

[17] Andreas Geiger et al. “Vision Meets Robotics: The KITTI Dataset”. In: *Int. J. Rob. Res.* Vol. 32. 11. 2013, pp. 1231–1237. DOI: 10.1177/0278364913491297.

[18] Xiuye Gu et al. “Open-vocabulary Object Detection via Vision and Language Knowledge Distillation”. In: *Int. Conf. Learn. Represent.* 2022. URL: https://doi.org/10.48550/arXiv.2104.13921.

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. “LVIS: A Dataset for Large Vocabulary Instance Segmentation”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 5356–5364. DOI: 10.1109/CVPR.2019.00550.

[20] Wooseok Ha and Rina Foygel Barber. “Alternating minimization and alternating descent over nonconvex sets”. In: *arXiv preprint arXiv:1709.04451* (2017). URL: https://arxiv.org/abs/1709.04451.

[21] Yulin He et al. “Recalling Unknowns Without Losing Precision: An Effective Solution to Large Model-Guided Open World Object Detection”. In: *IEEE Trans. Image Process.* Vol. 34. 2025. DOI: 10.1109/TIP.2024.3459589.

[22] Peng-Hao Hsu et al. “OpenM3D: Open Vocabulary Multi-view Indoor 3D Object Detection without Human Annotations”. In: *Int. Conf. Comput. Vis.* 2025. DOI: 10.1109/ICCV51701.2025.00813.

[23] Junjie Huang et al. “BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View”. In: *arXiv preprint arXiv:2112.11790* (2022). URL: https://arxiv.org/abs/2112.11790.

[24] Rui Huang et al. “Training an Open-Vocabulary Monocular 3D Object Detection Model without 3D Data”. In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. Vol. 37. 2024. DOI: 10.52202/079017-2303.

[25] Chao Jia et al. “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision”. In: *Int. Conf. Mach. Learn.* 2021. URL: https://proceedings.mlr.press/v139/jia21b.html.

[26] Alexander Kirillov et al. “Segment Anything”. In: *Int. Conf. Comput. Vis.* 2023, pp. 3992–4003. DOI: 10.1109/ICCV51070.2023.00371.

[27] Abhinav Kumar. “Generalizing Monocular 3D Object Detection”. In: *arXiv preprint arXiv:2508.19593* (2025). URL: https://arxiv.org/abs/2508.19593.

[28] Weicheng Kuo et al. “F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models”. In: *Int. Conf. Learn. Represent.* 2023. URL: https://arxiv.org/abs/2209.15639.

[29] Alex H. Lang et al. “PointPillars: Fast Encoders for Object Detection From Point Clouds”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 12689–12697. DOI: 10.1109/CVPR.2019.01298.

[30] Seungjae Lee, Hyungtae Lim, and Hyun Myung. "Patchwork++: Fast and Robust Ground Segmentation Solving Partial Under-Segmentation Using 3D Point Cloud". In: *Int. Conf. on Intelligent Robots and Systems*. 2022. DOI: 10.1109/IROS47612.2022.9981561.

[31] Fan Li et al. "TriCLIP-3D: A Unified Parameter-Efficient Framework for Tri-Modal 3D Visual Grounding based on CLIP". In: *ACM Int. Conf. Multimedia*. 2025, pp. 5257–5266. DOI: 10.1145/3746027.3755836.

[32] Feng Li et al. "Segment and Recognize Anything at Any Granularity". In: *Eur. Conf. Comput. Vis.* 2024, pp. 467–484. DOI: 10.1007/978-3-031-73195-2_27.

[33] Liunian Harold Li et al. "Grounded Language-Image Pre-training". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2022. DOI: 10.1109/CVPR52688.2022.01069.

[34] Yingwei Li et al. "DeepFusion: LiDAR-Camera Deep Fusion for Multi-Modal 3D Object Detection". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2022. DOI: 10.1109/CVPR52688.2022.01667.

[35] Zhiqi Li et al. "BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers". In: *Eur. Conf. Comput. Vis.* Springer Nature Switzerland, 2022. DOI: 10.1007/978-3-031-20077-9_1.

[36] Mingyu Liu et al. "A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook". In: *IEEE Trans. on Intelligent Vehicles*. Vol. 9. 11. 2024, pp. 7138–7164. DOI: 10.1109/TIV.2024.3394735.

[37] Qi Liu et al. "HQ-OV3D: A High Box Quality Open-World 3D Detection Framework based on Diffision Model". In: *arXiv preprint arXiv:2508.10935* (2025). URL: https://doi.org/10.48550/arXiv.2508.10935.

[38] Shilong Liu et al. "DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR". In: *Int. Conf. Learn. Represent.* 2022.

[39] Shilong Liu et al. "Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection". In: *Eur. Conf. Comput. Vis.* Springer-Verlag, 2024, pp. 38–55. DOI: 10.1007/978-3-031-72970-6_3.

[40] Taichi Liu et al. "Towards 3D Objectness Learning in an Open World". In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. Ed. by D. Belgrave et al. Curran Associates, Inc., 2025, pp. 91540–91566. DOI: 10.52202/085713-3063.

[41] Zhijian Liu et al. "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation". In: *IEEE Int. Conf. on Robotics and Automation*. 2023, pp. 2774–2781. DOI: 10.1109/ICRA48891.2023.10160968.

[42] Yuheng Lu et al. "Open-Vocabulary 3D Detection via Image-level Class and Debiased Cross-modal Contrastive Learning". In: *arXiv preprint arXiv:2207.01987* (2022). URL: https://arxiv.org/abs/2207.01987.

[43] Yuheng Lu et al. "Open-Vocabulary Point-Cloud Object Detection without 3D Annotation". In: *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE Computer Society, 2023, pp. 1190–1199. DOI: 10.1109/CVPR52729.2023.00121.

[44] Claudia Malzer and Marcus Baum. "HDBSCAN($\hat{\varepsilon}$): An Alternative Cluster Extraction Method for HDBSCAN". In: *CoRR* abs/1911.02282 (2019). URL: https://doi.org/10.48550/arXiv.1911.02282.

[45] Jiageng Mao et al. “3D Object Detection for Autonomous Driving: A Comprehensive Survey”. In: *Int. J. Comput. Vis.* Vol. 131. 2023, pp. 1909–1963. DOI: 10.1007/s11263-023-01790-1.

[46] Jiayuan Mao, Min Niu, Zhitong Bai, et al. “One Million Scenes for Autonomous Driving: ONCE Dataset”. In: *NeurIPS Datasets and Benchmarks*. 2021, pp. 12521–12530. URL: https://doi.org/10.48550/arXiv.2106.11037.

[47] Arsalan Mousavian et al. “3D Bounding Box Estimation Using Deep Learning and Geometry”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 5632–5640. DOI: 10.1109/CVPR.2017.597.

[48] Mahyar Najibi et al. “Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving”. In: *Int. Conf. Comput. Vis.* 2023, pp. 8602–8612. DOI: 10.1109/ICCV51070.2023.00790.

[49] OpenAI. “GPT-4 Technical Report”. In: *arXiv preprint arXiv:2303.08774* (2024). URL: https://doi.org/10.48550/arXiv.2303.08774.

[50] Maxime Oquab et al. “DINOv2: Learning Robust Visual Features without Supervision”. In: *arXiv preprint arXiv:2304.07193* (2023). URL: https://doi.org/10.48550/arXiv.2304.07193.

[51] Charles R. Qi et al. “Frustum PointNets for 3D Object Detection From RGB-D Data”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018. DOI: 10.1109/CVPR.2018.00102.

[52] Charles R. Qi et al. “PointNet++: deep hierarchical feature learning on point sets in a metric space”. In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. 2017, pp. 5105–5114.

[53] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Int. Conf. Mach. Learn.* (2021). Ed. by Marina Meila and Tong Zhang, pp. 8748–8763.

[54] Yongming Rao et al. “DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2022. DOI: 10.1109/CVPR52688.2022.01755.

[55] Tianhe Ren, Shilong Liu, Hao Zhang, et al. “Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks”. In: *arXiv preprint arXiv:2401.14159* (2024). URL: https://doi.org/10.48550/arXiv.2401.14159.

[56] Olga Russakovsky et al. “ImageNet Large Scale Visual Recognition Challenge”. In: *Int. J. Comput. Vis.* Vol. 115. 3. 2015, pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[57] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. “PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 770–779. DOI: 10.1109/CVPR.2019.00086.

[58] Poulami Sinhamahapatra et al. “Zero-Shot Open-Vocabulary OOD Object Detection and Grounding using Vision Language Models”. In: *Northern Lights Deep Learning Conf.* Vol. 265. 2025, pp. 230–238. URL: https://proceedings.mlr.press/v265/sinhamahapatra25a.html.

[59] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. “SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2015, pp. 567–576. DOI: 10.1109/CVPR.2015.7298655.

[60] Pei Sun et al. “Scalability in Perception for Autonomous Driving: Waymo Open Dataset”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020, pp. 2446–2454. DOI: 10.1109/CVPR42600.2020.00252.

[61] Zhihao Sun et al. “OV-GT3D: A generalizable open-vocabulary two-stage 3D detector with dual path distillation”. In: *Pattern Recognition*. Vol. 171. 2026, p. 112156. DOI: 10.1016/j.patcog.2025.112156.

[62] Ivan E. Sutherland and Gary W. Hodgman. “Reentrant polygon clipping”. In: *Commun. ACM* 17.1 (1974), pp. 32–42. DOI: 10.1145/360767.360802.

[63] Noor Ul Ain Tahir et al. “Object Detection in Autonomous Vehicles under Adverse Weather: A Review of Traditional and Deep Learning Approaches”. In: *Algorithms* 17.3 (2024), p. 103. URL: https://doi.org/10.3390/a17030103.

[64] Ashish Vaswani et al. “Attention is All You Need”. In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. Curran Associates, Inc., 2017, pp. 6000–6010.

[65] Ke Wang et al. “Performance and Challenges of 3D Object Detection Methods in Complex Scenes for Autonomous Driving”. In: *IEEE Trans. on Intelligent Vehicles*. Vol. 8. 2. 2023, pp. 1699–1716. DOI: 10.1109/TIV.2022.3213796.

[66] Qitai Wang et al. “Immortal tracker: Tracklet never dies”. In: *arXiv preprint arXiv:2111.13672* (2021). URL: https://doi.org/10.48550/arXiv.2111.13672.

[67] Yue Wang et al. “DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries”. In: *arXiv preprint arXiv:2110.06922* (2021). URL: https://arxiv.org/abs/2110.06922.

[68] Zhenyu Wang et al. “OV-Uni3DETR: Towards Unified Open-Vocabulary 3D Object Detection via Cycle-Modality Propagation”. In: *Eur. Conf. Comput. Vis.* 2024. DOI: 10.1007/978-3-031-72970-6_5.

[69] Benjamin Wilson, Jiasen Qi, Saurabh Agarwal, et al. “Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting”. In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4734ba6f3de83d861c3176a6273cac6d-Paper-round2.pdf.

[70] Zhongyu Xia et al. “OpenAD: Open-World Autonomous Driving Benchmark for 3D Object Detection”. In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. 2025. DOI: 10.52202/085713-4084.

[71] Yan Yan, Yuxing Mao, and Bo Li. “SECOND: Sparsely Embedded Convolutional Detection”. In: *Sensors* 18.10 (2018), p. 3337. DOI: 10.3390/s18103337.

[72] Timing Yang, Yuanliang Ju, and Li Yi. “ImOV3D: Learning Open-Vocabulary Point Clouds 3D Object Detection from Only 2D Images”. In: *Adv. Neural Inform. Process. Syst. (NeurIPS)*. 2024. DOI: 10.5555/3737916.3742402.

[73] Jin Yao et al. “Open Vocabulary Monocular 3D Object Detection”. In: *Int. Conf. on 3D Vision*. 2026. DOI: 10.1109/3DV69130.2026.00171.

[74] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. “Center-Based 3D Object Detection and Tracking”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2021, pp. 11784–11793. DOI: 10.1109/CVPR46437.2021.01161.

[75] Fisher Yu et al. “BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020, pp. 2636–2645. DOI: 10.1109/CVPR42600.2020.00271.

[76] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. “Open-Vocabulary Object Detection Using Captions”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2021. DOI: 10.1109/CVPR46437.2021.01416.

[77] Hanxue Zhang et al. “Detect Anything 3D in the Wild”. In: *Int. Conf. Comput. Vis.* 2025. DOI: 10.1109/ICCV51701.2025.00480.

[78] Haomeng Zhang et al. “Auto-Vocabulary 3D Object Detection”. In: *arXiv preprint arXiv:2512.16077* (2025). URL: https://doi.org/10.48550/arXiv.2512.16077.

[79] Hu Zhang et al. “OpenSight: A Simple Open-Vocabulary Framework for LiDAR-Based Object Detection”. In: *Eur. Conf. Comput. Vis.* Springer-Verlag, 2024, pp. 1–19. DOI: 10.1007/978-3-031-72907-2_1.

[80] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Adding Conditional Control to Text-to-Image Diffusion Models”. In: *Int. Conf. Comput. Vis.* 2023. DOI: 10.1109/ICCV51070.2023.00355.

[81] Renrui Zhang et al. “MonoDETR: Depth-guided Transformer for Monocular 3D Object Detection”. In: *Int. Conf. Comput. Vis.* 2023. DOI: 10.1109/ICCV51070.2023.00840.

[82] Yiwu Zhong et al. “RegionCLIP: Region-based Language-Image Pretraining”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2022. DOI: 10.1109/CVPR52688.2022.01629.

[83] Xingyi Zhou et al. “Detecting Twenty-thousand Classes using Image-level Supervision”. In: *Eur. Conf. Comput. Vis.* 2022. DOI: 10.1007/978-3-031-20077-9_21.

[84] Yin Zhou and Oncel Tuzel. “VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 4490–4499. DOI: 10.1109/CVPR.2018.00472.

Appendix A

Systematic Literature Review

This appendix details the systematic literature review process conducted to support the State of the Art (Chapter 2). The methodology adopts principles from the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) framework to ensure a transparent, reproducible, and structured search and selection process.

A.1 Search Queries and Initial Results

The search strategy was executed across four major academic databases (IEEE Xplore, ACM Digital Library, Scopus, and arXiv) using queries targeting Title, Abstract, and Keywords (with database-specific syntax adapters for boolean negation). To isolate outdoor autonomous driving perception, negative filter strings were applied across all queries to exclude unrelated domains (such as medical imaging, satellite/aerial remote sensing, adversarial attacks, temporal tracking, and V2X infrastructure).

The specific queries were defined as follows:

Q1 (Closed-Set 3D Detection): Focuses on traditional 3D Object Detection (OD) within the context of Autonomous Driving (AD) while explicitly excluding open-vocabulary terms:

```
("3D") AND ("object detection") AND ("camera" OR "lidar" OR "point
cloud") AND ("autonomous driving") NOT ("open-vocabulary" OR "open-world"
OR "prediction" OR "medical" OR "satellite" OR "radar" OR "attack"
OR "video" OR "roadside" OR "V2X" OR "temporal" OR "collaborative
perception" OR "active learning" OR "tracking")
```

Q2 (General Open-Vocabulary): Captures Open-Vocabulary Object Detection (OV-OD) methods across general domains to identify foundational vision-language mechanisms:

```
("object detection") AND ("camera" OR "image" OR "lidar" OR "point
cloud") AND ("open-vocabulary" OR "open-world") NOT ("aerial" OR
"UAV" OR "medical" OR "satellite" OR "radar" OR "attack" OR "video"
OR "roadside" OR "V2X" OR "temporal" OR "collaborative perception"
OR "active learning" OR "tracking")
```

Q3 (3D Open-Vocabulary for AD): Targets the intersection of 3D Open-Vocabulary (OV) methods for AD. By not restricting this query strictly to object detection, it allows capturing broader spatial tasks informing 3D object discovery:

```
("3D") AND ("open-vocabulary" OR "open-world") AND ("autonomous
driving" OR "autonomous vehicle" OR "kitti" OR "nuscenes" OR "waymo")
NOT ("medical" OR "satellite" OR "radar" OR "attack" OR "video"
OR "roadside" OR "V2X" OR "temporal" OR "collaborative perception"
OR "active learning" OR "tracking")
```

The initial database search yielded a total of 1,990 records. The raw distribution of these results per query and database is detailed in Table A.1.

Table A.1: Initial Search Results prior to filtering.

Query	ACM	arXiv	IEEE	Scopus	Total
Q1	37	227	412	676	**1,352**
Q2	61	172	65	235	**533**
Q3	10	46	20	29	**105**
Total	**108**	**445**	**497**	**940**	**1,990**

A.2 Screening and Exclusion Criteria

To reduce the identified records ($n = 1,990$) to a relevant set, specific exclusion criteria were applied in two stages, as visualised in Figure A.1.

A.2.1 Automatic Filtering (Pre-Screening)

Before manual screening, temporal and citation filters were applied to ensure relevance and quality:

- **E1 (Applied to Q1):** Records published before 2023 were excluded ($n = 412$) to ensure the baseline comparison is restricted to the current state-of-the-art.
- **E2 (Applied to Q1):** Records published between 2023 and 2024 with fewer than 20 citations, or between 2025 and 2026 with fewer than 3 citations, were excluded ($n = 722$).
- **E3 (Applied to Q2/Q3):** Records published before 2021 were excluded ($n = 21$) to focus on the modern, post-CLIP paradigm of OV perception.
- **E4 (Applied to Q2):** Records published between 2021 and 2024 with fewer than 20 citations were excluded ($n = 138$).

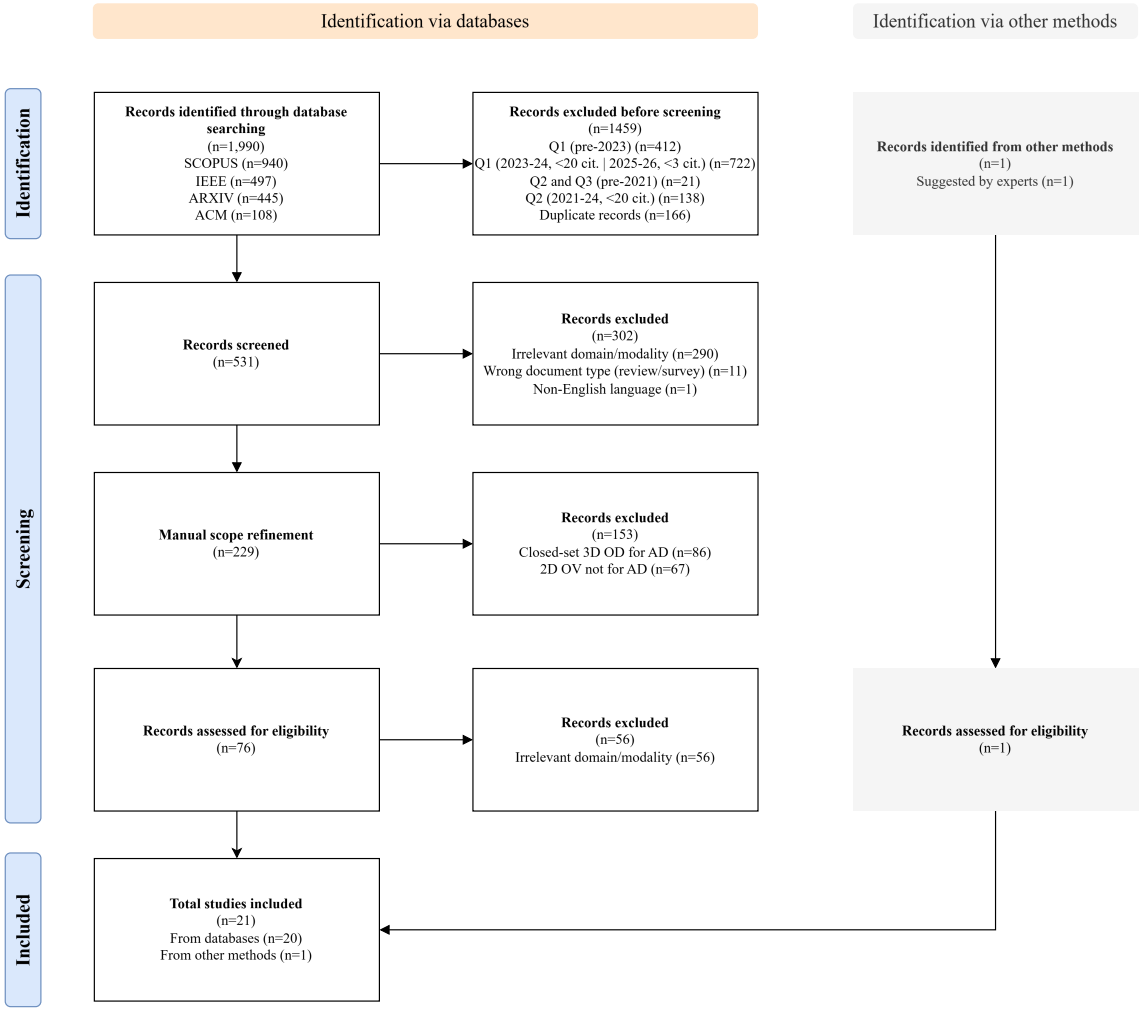


Figure A.1: PRISMA 2020 Flow Diagram detailing the identification, screening, and inclusion phases.

A.2.2 Manual Screening and Eligibility

Records passing the automatic filters were screened based on title and abstract, followed by a full-text eligibility assessment. The following manual exclusion criteria were applied:

- **E5:** Removal of duplicated records ($n = 166$).
- **E6:** Exclusion of papers focusing on irrelevant domains or incorrect modalities ($n = 346$).
- **E7:** Exclusion of reviews, surveys, abstracts, or tutorials ($n = 11$).
- **E8:** Exclusion of non-English papers ($n = 1$).
- **E9:** Narrowing the qualitative focus to OV-OD within AD and related 3D domains. While the excluded closed-set 3D OD and general 2D OV works are retained as contextual references to establish the state-of-the-art baseline, they are omitted from the deep qualitative synthesis. This refinement allows for a more in-depth analysis of the specific architectures most relevant to the thesis objectives, managing the high volume of literature while maintaining the thematic focus ($n = 153$).

Following this screening process, a total of 21 studies were included for qualitative synthesis presented in Chapter 2.

Appendix B

Baseline Reproduction

B.1 Structure of the OV-SCAN Reference Labels

To validate the reproduction of the Global HDBSCAN Pipeline, the following files comprising the OV-SCAN Reference Labels were analysed:

- `nuscenes_infos_10sweeps_train_ov_task.pkl`: Contains 28,130 entries corresponding to each frame, with information for 9 sweeps to merge with the keyframe.
- `nuscenes_dbinfos_10sweeps_withvelo_full_ov.pkl`: A dictionary containing paths to point cloud crops, alongside metadata storing the original 2D bounding boxes, their associated text labels, and confidence scores.
- `train.zip`: Contains object embeddings extracted using DFN5B-CLIP-ViT-H-14.

Across the provided database, a total of 196,090 visual crops were successfully extracted and deemed fit for alignment loss across the dataset. Additionally, the camera view distribution was: CAM_BACK (30.46%), CAM_FRONT (19.39%), CAM_FRONT_LEFT (13.99%), CAM_BACK_LEFT (13.15%), CAM_FRONT_RIGHT (12.82%), and CAM_BACK_RIGHT (10.19%).

B.2 Subset Selected to Preserve Class Distribution

Training the OV-SCAN Baseline on the complete nuScenes dataset under the Full Benchmark Protocol requires approximately two days of continuous computation on available hardware. To mitigate this significant bottleneck during iterative hyperparameter tuning and component ablations, we used a greedy selection algorithm to extract a reduced subset of the original scenes while preserving the class distribution of the global dataset.

The algorithm begins by computing a summary vector representing the class composition of each scene. For a given scene i, let F_i be the set of its frames (sweeps). For each frame $j \in F_i$, we compute a class count vector $v_{i,j} \in \mathbb{R}^{|C|}$, where $v_{i,j}(c)$ represents the number of objects of class c in

that specific frame. The total class count vector for the entire scene i, denoted v_i, is the sum of its frame-level vectors:

$$v_i = \sum_{j=1}^{|F_i|} v_{i,j} \tag{B.1}$$

We then define the target class distribution for the entire training set T. First, we compute the global count vector V_T by summing all scene vectors:

$$V_T = \sum_{i=1}^{|T|} v_i \tag{B.2}$$

This vector V_T is normalised by its L1 norm to produce the target distribution vector p_T, where each element $p_T(c)$ represents the global frequency of class c:

$$p_T = \frac{V_T}{\|V_T\|_1} \tag{B.3}$$

The selection process initialises an empty subset $S_0 = \emptyset$ and iteratively adds one scene at a time for k steps. At each step k, the algorithm evaluates every remaining candidate scene $s' \in T \setminus S_k$. We define $V_{\text{candidate}}$ as the hypothetical total class count vector if s' were added:

$$V_{\text{candidate}} = \left(\sum_{i=1}^{|S_k|} v_i \right) + v_{s'} \tag{B.4}$$

where $v_{s'}$ represents the class count vector of the candidate scene. This candidate vector is normalised to yield the hypothetical class distribution $p_{\text{candidate}}$:

$$p_{\text{candidate}} = \frac{V_{\text{candidate}}}{\|V_{\text{candidate}}\|_1} \tag{B.5}$$

The algorithm then greedily selects the scene s^* that minimises the L1 distance between the resulting candidate distribution and the global target distribution p_T:

$$s^* = \underset{s' \in T \setminus S_k}{\operatorname{argmin}} \|p_{\text{candidate}} - p_T\|_1 \tag{B.6}$$

The chosen scene s^* is added to the subset ($S_{k+1} = S_k \cup \{s^*\}$), and the process repeats until the target subset size is reached.

To validate the stability of this greedy selection algorithm, we forced it to generate stratified subsets across varying target scene volumes. Because the algorithm relies on a greedy heuristic, changing the target volume forces it to explore entirely different combinations of scenes to satisfy the global class distribution. The OV-SCAN Baseline was trained on these varying subsets using the OV-SCAN Reference Labels, and the results are presented in Table B.1.

As demonstrated in Table B.1, despite exploring different scene combinations, the greedy algorithm consistently maintains distributional alignment with the global dataset (L1 Distance

Table B.1: Performance of the OV-SCAN Baseline trained across stratified subsets of varying sizes. All models were supervised using the OV-SCAN Reference Labels. The consistently low distributional distances prove the stability of the greedy selection algorithm. The 30% subset was selected as it provides sufficient object volume to establish a stable detection baseline while remaining computationally efficient.

Scenes (%)	**Objects (%)**	**L1 Dist** ($\times 10^2$)	**L∞ Dist** ($\times 10^2$)	**mAP (%)**
15	5.70	0.99	**0.22**	16.43
20	8.70	0.97	0.41	19.14
25	12.08	**0.77**	0.30	21.68
30	15.36	0.82	**0.22**	**23.60**

remains $\leq 0.99 \times 10^{-2}$ across all runs). This validates the robustness of our subset generation methodology.

With the selection algorithm validated, the 30% scene subset was chosen as the optimal configuration for rapid prototyping. As expected, detection performance naturally scales with the raw volume of objects extracted. By capturing 15.36% of the total dataset objects, the 30% subset achieves a baseline of 23.60% mAP. This provides a performance threshold stable enough to reliably measure downstream architectural ablations without being overwhelmed by dataset noise, while remaining computationally small enough to allow for rapid iteration.

B.3 Extended Grid Search Results for Proposal Generation Reproduction

Table B.2 presents the complete results of the hyperparameter grid search conducted for the Global HDBSCAN Pipeline, as discussed in Section 4.4.2. To isolate these tuning metrics from downstream architectural variance, all combinations were evaluated under the Representative Subset Protocol. The grid search evaluated combinations across three modules:

1. **HDBSCAN** (*min_cluster_size, min_samples,* ε)
2. **Cross-Modal Association (CMA)** (τ_{match}, N_{matched})
3. **Post-Processing (PP)** (3D Non-Maximum Suppression (NMS) Intersection-over-Union (IoU) threshold)

All metrics are presented as percentages (%) evaluated at an IoU threshold of ≥ 0.5. The table is sorted by the F1-Score of our generated proposals against the nuScenes Ground Truth.

Table B.2: Comprehensive Grid Search Results evaluated under the Representative Subset Protocol. Metrics are presented as percentages (%).

Parameters			Ours vs. Ground Truth			OV-SCAN Ref. Labels vs. Ground Truth			Ours vs. OV-SCAN Ref. Labels		
HDBSCAN	CMA	PP Thresh	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
(5, 3, 0.8)	(0.5m, 1)	0.5	13.0	27.8	**17.8**	13.3	50.4	21.0	16.0	28.5	20.5
(5, 3, 0.8)	(1.5m, 3)	0.5	13.3	25.3	17.5	13.3	50.4	21.0	14.0	28.0	18.7
(5, 3, 0.8)	(0.5m, 3)	0.5	12.9	27.1	17.5	13.3	50.4	21.0	15.5	28.1	20.0
(5, 3, 0.8)	(0.5m, 1)	0.2	12.5	28.7	17.4	13.3	50.4	21.0	16.6	27.5	20.7
(5, 3, 0.8)	(1.5m, 3)	0.2	13.0	26.3	17.4	13.3	50.4	21.0	14.6	27.4	19.0
(5, 3, 0.8)	(0.5m, 3)	0.2	12.6	28.1	17.4	13.3	50.4	21.0	16.1	27.4	20.3
(5, 3, 0.5)	(1.5m, 3)	0.2	12.9	25.8	17.2	13.3	50.4	21.0	14.2	27.0	18.6
(5, 3, 0.5)	(1.5m, 3)	0.5	13.1	24.7	17.1	13.3	50.4	21.0	13.6	27.4	18.2
(5, 3, 0.5)	(0.5m, 1)	0.5	12.6	26.5	17.0	13.3	50.4	21.0	15.3	27.5	19.7
(5, 3, 0.5)	(0.5m, 3)	0.2	12.3	27.2	16.9	13.3	50.4	21.0	15.7	27.0	19.9
(5, 3, 0.5)	(0.5m, 3)	0.5	12.5	26.1	16.9	13.3	50.4	21.0	15.2	27.5	19.5
(5, 3, 0.5)	(0.5m, 1)	0.2	12.1	27.4	16.8	13.3	50.4	21.0	15.9	26.7	20.0
(5, 3, 0.8)	(0.5m, ∞)	0.5	12.3	25.7	16.7	13.3	50.4	21.0	14.9	27.1	19.2
(5, 3, 0.8)	(0.5m, ∞)	0.2	12.1	26.6	16.6	13.3	50.4	21.0	15.3	26.5	19.4
(5, 3, 0.5)	(0.5m, ∞)	0.2	11.8	25.9	16.2	13.3	50.4	21.0	15.0	26.1	19.1
(5, 3, 0.5)	(0.5m, ∞)	0.5	12.0	24.9	16.2	13.3	50.4	21.0	14.5	26.4	18.7
(20, 10, 0.5)	(0.5m, 1)	0.5	10.8	27.9	15.6	13.3	50.4	21.0	16.0	23.6	19.1
(20, 10, 0.5)	(0.5m, 3)	0.5	10.6	27.1	15.3	13.3	50.4	21.0	15.8	23.5	18.9
(20, 10, 0.5)	(0.5m, 1)	0.2	10.4	29.0	15.3	13.3	50.4	21.0	16.8	22.8	19.3
(20, 10, 0.5)	(0.5m, 3)	0.2	10.3	28.2	15.1	13.3	50.4	21.0	16.4	22.8	19.1
(20, 10, 0.5)	(1.5m, 3)	0.5	10.8	24.4	14.9	13.3	50.4	21.0	14.0	23.5	17.5
(20, 10, 0.5)	(0.5m, ∞)	0.5	10.4	26.3	14.9	13.3	50.4	21.0	15.5	23.3	18.7
(20, 10, 0.5)	(1.5m, 3)	0.2	10.4	25.5	14.8	13.3	50.4	21.0	14.6	22.6	17.8
(20, 10, 0.5)	(0.5m, ∞)	0.2	10.1	27.4	14.7	13.3	50.4	21.0	16.2	22.6	18.9
(5, 30, 0.5)	(0.5m, 1)	0.5	8.9	25.5	13.2	13.3	50.4	21.0	14.6	19.4	16.7
(5, 30, 0.5)	(0.5m, 1)	0.2	8.5	26.3	12.8	13.3	50.4	21.0	15.3	18.7	16.9
(5, 30, 0.5)	(0.5m, 3)	0.5	8.7	24.6	12.8	13.3	50.4	21.0	14.5	19.4	16.6
(5, 30, 0.5)	(0.5m, ∞)	0.5	8.5	24.1	12.6	13.3	50.4	21.0	14.3	19.3	16.5
(5, 30, 0.5)	(0.5m, 3)	0.2	8.4	25.5	12.6	13.3	50.4	21.0	15.1	18.8	16.7
(5, 30, 0.5)	(1.5m, 3)	0.5	8.8	21.9	12.6	13.3	50.4	21.0	12.8	19.7	15.5
(5, 30, 0.5)	(1.5m, 3)	0.2	8.5	22.9	12.4	13.3	50.4	21.0	13.5	19.0	15.8
(5, 30, 0.5)	(0.5m, ∞)	0.2	8.2	25.0	12.4	13.3	50.4	21.0	14.9	18.7	16.6

Appendix C

Implementation Hyperparameters

To ensure full reproducibility, the exhaustive hyperparameter configurations used throughout the offline pseudo-label generation pipeline and the online 3D detection training phase are detailed in the tables below.

C.1 Pseudo-Label Generation Pipeline

Table C.1: Hyperparameters for 2D Proposal Generation (Grounding DINO) and Cross-Modal Association.

Module	Parameter	Value
2D Detections	Box Confidence Threshold	0.35
	Text Confidence Threshold	0.35
	2D NMS Threshold	0.60
Cross-Modal	Valid Depth Range (m)	$[2.0, 60.0]$
	Point-to-Ray Matching Threshold (τ_{match})	0.5m
	Max Matched Clusters per Frustum	1
	Global 3D NMS Threshold	0.50

Table C.2: Point Cloud Clustering configurations. The Global HDBSCAN Pipeline uses global HDBSCAN, whereas the proposed Local DBSCAN Pipeline replaces it with frustum-localised DBSCAN.

Parameter	Global HDBSCAN Pipeline	Local DBSCAN Pipeline
Point Cloud Range	$[-51.2, -51.2, -5.0] \rightarrow [51.2, 51.2, 3.0]$	
`min_cluster_size`	5	N/A
`min_samples`	3	4
ε (Epsilon)	0.8	0.8
Metric	Euclidean	Euclidean

Table C.3: Particle Swarm Optimisation (PSO) hyperparameters used for offline 3D bounding box fitting.

Parameter	**Description**	**Value**
N_{swarm}	Number of particles	50
N_{iter}	Optimisation iterations	500
$w_{\text{init}} \rightarrow w_{\text{end}}$	Cosine decayed inertia weight	$10.0 \rightarrow 0.1$
c_1, c_2	Cognitive and Social coefficients	1.0, 1.0
C_{noise}	Initialisation spatial noise	0.1
λ_1	Density Cost Weight	5.0
λ_2	L-Shape Cost Weight	1.0
λ_3	Surface Proximity Cost Weight	1.0
γ	2D Alignment Cost Weight	3.0
C_{surface}	Surface proximity capping constant	5.0m

C.2 OV-D3D Training Configuration

Table C.4: Loss weighting coefficients and network architecture parameters for the OV-D3D model.

Component	**Parameter**	**Value**
Loss Weights	Heatmap Objectness (λ_{hm})	1.0
	Dense BEV Dimensions (λ_{dim})	1.0
	Bounding Box Regression (λ_{box})	0.25
	Objectness (λ_{cls})	0.25
	VLM Alignment (λ_{align})	1.0
Transformer	Hidden Channels (D_{hidden})	128
	Attention Heads	8
	FFN Channels	256
	Dropout	0.1
VLM Alignment	Target Latent Space	CLIP ViT-H-14 (1024-D)

Appendix D

Extended Metrics and Semantic Analysis

D.1 Full Class-Agnostic Recall Results

Table D.1 presents the exhaustive class-agnostic recall evaluated at a standard 2.0 m centre-distance threshold across all nuScenes categories.

Table D.1: Full Class-Agnostic Recall (%) at 2.0m across all nuScenes categories. Held-out zero-shot categories are marked with (†). OV-D3D preserves high proposal recall across novel geometries, whereas class-coupled proposals (the OV-SCAN Baseline) suffer recall drops.

		OV-D3D (Ours)		**1-Channel Control**		**OV-SCAN Baseline**	
Macro Class	**GT Count**	Standard	Holdout	Standard	Holdout	Standard	Holdout
Car	59,486	83.67	83.42	83.95	83.63	75.96	75.86
Pedestrian	29,984	76.60	77.28	76.43	76.85	83.59	83.74
Barrier†	21,877	77.85	**65.61**	77.28	67.69	70.23	41.71
Truck	13,099	67.98	68.61	68.40	68.47	63.67	63.91
Traffic Cone	12,074	85.68	88.55	85.21	88.68	87.40	85.93
Trailer†	3,546	34.07	**30.77**	32.57	34.29	25.72	23.01
Bus	2,896	59.98	56.28	63.05	58.91	58.56	59.29
Construction Vehicle	2,301	62.41	61.80	62.15	63.32	48.72	48.85
Motorcycle†	2,180	88.17	**87.61**	89.50	87.48	81.70	77.66
Bicycle	2,037	87.73	87.19	88.46	88.56	81.89	79.14
Global Recall (2.0 m)	**149,480**	**78.35**	**76.70**	78.39	77.18	74.72	70.29

To evaluate localisation across stricter and coarser thresholds, Table D.2 reports the mean Agnostic Recall (mAR) averaged over $\mathbb{D} = \{0.5, 1.0, 2.0, 4.0\}$ m.

D.2 Full Semantic Confusion Matrices

As discussed in Section 4.3.5, diagnosing the cross-modal alignment bottleneck requires isolating the continuous semantic embedding space from discrete detection decisions. The matrices presented in this appendix illustrate the cross-modal similarity distributions across the 10 nuScenes classes

Table D.2: Class-Agnostic Recall across distance thresholds ($\mathbb{D} = \{0.5, 1.0, 2.0, 4.0\}$ m) and mean Agnostic Recall (mAR) evaluated globally and on novel holdout classes.

Evaluation Scope	Model Split	AR@0.5m	AR@1.0m	AR@2.0m	AR@4.0m	mAR
Global Dataset	OV-SCAN Baseline (Standard)	56.09	66.45	74.72	80.02	69.32
	OV-SCAN Baseline (Holdout)	52.63	61.74	70.29	76.60	65.31
	1-Channel Control (Standard)	51.46	65.21	78.39	86.12	70.29
	1-Channel Control (Holdout)	49.40	63.14	77.18	85.01	68.68
	OV-D3D (Standard)	51.78	65.55	78.35	86.04	70.43
	OV-D3D (Holdout, Ours)	49.53	62.93	**76.70**	**84.67**	**68.46**
Novel: Barrier	OV-SCAN Baseline (Standard)	36.99	56.83	70.23	78.48	60.63
	OV-SCAN Baseline (Holdout)	13.93	26.05	41.71	56.90	34.64
	1-Channel Control (Holdout)	12.34	39.31	67.69	80.28	49.91
	OV-D3D (Holdout, Ours)	12.30	**38.70**	**65.61**	**78.50**	**48.78**
Novel: Motorcycle	OV-SCAN Baseline (Standard)	64.82	75.05	81.70	86.65	77.05
	OV-SCAN Baseline (Holdout)	58.49	67.89	77.66	85.32	72.34
	1-Channel Control (Holdout)	55.60	76.47	87.48	92.43	77.99
	OV-D3D (Holdout, Ours)	55.00	**75.32**	**87.61**	**92.39**	**77.58**
Novel: Trailer	OV-SCAN Baseline (Standard)	9.73	18.22	25.72	45.54	24.80
	OV-SCAN Baseline (Holdout)	6.91	14.18	23.01	39.14	20.81
	1-Channel Control (Holdout)	6.68	17.60	34.29	54.82	28.35
	OV-D3D (Holdout, Ours)	5.47	**14.64**	**30.77**	**53.36**	**26.06**

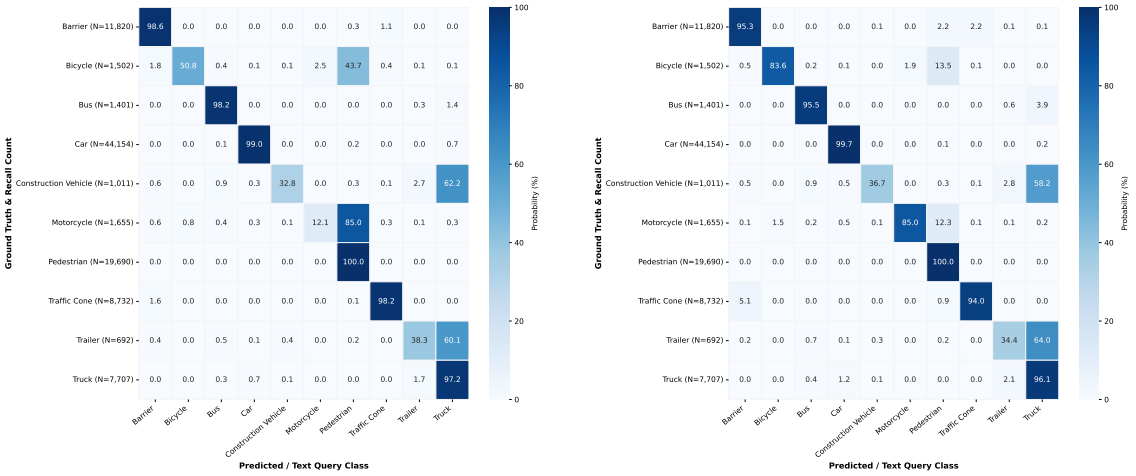
for both the OV-SCAN Baseline and our proposed OV-D3D architecture, evaluated under both the Full Benchmark Protocol and the Zero-Shot Holdout Protocol.

Mathematical Definition and Interpretation. Crucially, these figures do not represent standard discrete classification confusion matrices based on argmax predictions or detection counts. Instead, each row represents the mean probability distribution assigned across the textual vocabulary for all proposals matched to a specific ground-truth class. For a given ground-truth object i, the model produces an unscaled cosine similarity vector $\mathbf{s}_i \in \mathbb{R}^{10}$ against the text embeddings. In the raw matrices (Figure D.2), the cells report the mean cosine similarity $\bar{s}_{i,j}$. In the softmax matrices (Figure D.1 and Figure D.3), the cells display the row-normalised mean probability after temperature scaling:

$$P(c = j \mid \mathbf{s}_i) = \frac{\exp(s_{i,j} \cdot \tau)}{\sum_{k=1}^{10} \exp(s_{i,k} \cdot \tau)}, \quad \text{with } \tau = 100. \tag{D.1}$$

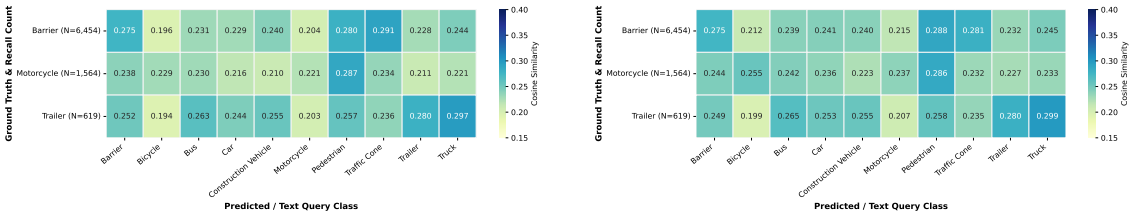
Relationship Between Probability Mass and 0 mAP. The non-zero diagonal entries for novel categories in Figure D.3 (e.g., 14.9% for `Barrier` and 12.4% for `Trailer` in OV-D3D) reflect average assigned softmax probabilities across matched ground-truth proposals, rather than discrete argmax classification counts. In standard 3D detection benchmarks, evaluation relies on top-1 class scoring. On individual instances, visually proximate base classes consistently achieve marginally higher raw similarities (e.g., "truck" at 0.298 vs. "trailer" at 0.285), systematically outranking the novel class. Because novel categories rarely secure the top rank, true-positive detections

remain sufficiently scarce that precision and recall stay below the official 0.1 nuScenes benchmark evaluation cutoff, resulting in an official score of 0 mAP.



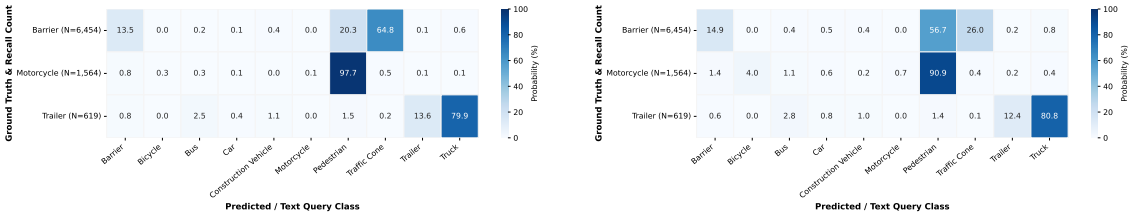
(a) OV-SCAN Baseline (b) OV-D3D (Ours)

Figure D.1: Complete 10×10 semantic confusion matrices representing **Softmax probabilities** ($\tau = 100$) on the Full Benchmark Protocol. The temperature multiplier forces confident predictions, often exaggerating marginal raw similarity differences.



(a) OV-SCAN Baseline (b) OV-D3D (Ours)

Figure D.2: Semantic confusion matrices representing **Raw Cosine similarities evaluated under the Zero-Shot Holdout Protocol**. Even without the text prior, the raw embeddings show moderate structural recognition before Softmax scaling is applied.



(a) OV-SCAN Baseline (b) OV-D3D (Ours)

Figure D.3: Semantic confusion matrices representing **Softmax probabilities evaluated on the Zero-Shot Holdout Protocol**. Both methods overwhelmingly default their predictions to structurally related base classes when deprived of the text prior during training.

Appendix E

Preliminary Explorations and Paradigm Limits

This appendix details preliminary explorations conducted to push the theoretical boundaries of the offline pseudo-labelling pipeline. Specifically, we attempted to remove two fundamental constraints: single-frame independence and categorical dimensional anchors.

Disclaimer: The quantitative metrics reported in this appendix were generated on a modified experimental configuration using the Representative Subset Protocol. Due to these pipeline alterations, the absolute Mean Average Precision (mAP) and IoU values are used strictly for internal comparative diagnosis within these preliminary tests and are not directly comparable to the benchmarked metrics in Chapter 4.

E.1 Temporal Tracklet Optimisation

To evaluate whether temporally aggregating pseudo-labels could improve geometric stability, we explored linking 2D proposals across consecutive frames prior to 3D lifting. We employed a Simple Online and Realtime Tracking (SORT) [4] algorithm, using a constant-velocity Kalman Filter to associate 2D proposals.

The objective was to transition the PSO from a single-frame process (optimising 7 variables per detection) to a global tracklet process (optimising $7 \times N_{\text{frames}}$ variables). Because optimising all variables simultaneously across long tracklets proved mathematically unstable, we developed an Alternating Optimisation PSO (AO-PSO) [20]. Assuming that the physical size of a rigid object remains constant, we locked the dimensional variables (w, l, h) globally across the entire tracklet, while iteratively optimising the positional variables (x, y, z, θ) on a per-frame basis.

For subsets of highly visible objects, locking the dimensions globally proved effective at stabilising the bounding box geometry against frame-by-frame occlusion (as qualitatively visualised in Figure E.1).

However, this global optimisation proved extremely sensitive to tracklet purity. The reliance on basic 2D tracking heuristics led to frequent errors in the 2D associations, particularly identity

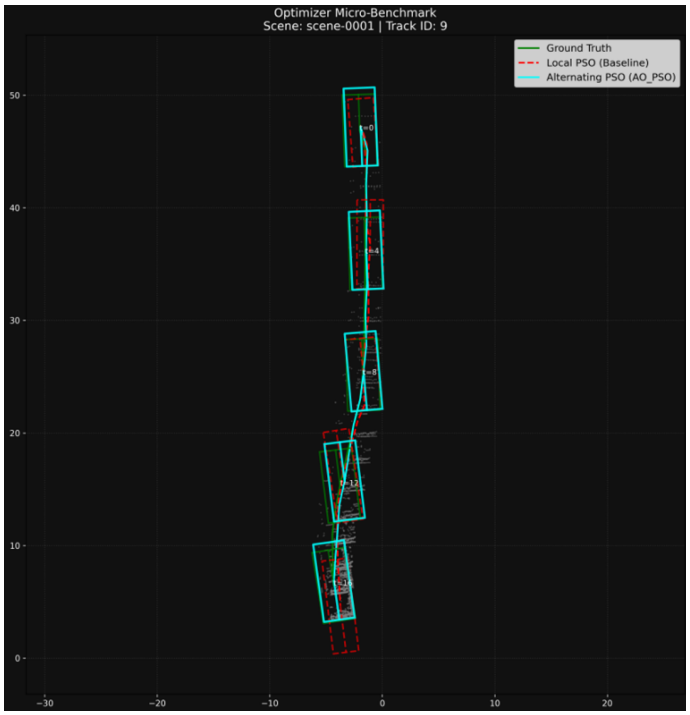


Figure E.1: Comparative Bird's Eye View (BEV) trajectory over 5 frames. Compared to independent single-frame optimisation (which fluctuates in size due to occlusion), the AO-PSO successfully stabilises the global dimensions of the vehicle across the tracklet.

switches during occlusion. When an impure tracklet is fed into the AO-PSO, the forced global dimensions propagate geometric noise across all incorrectly associated frames. Counter-intuitively, training the downstream 3D detectors (the OV-SCAN Baseline and OV-D3D) on these temporal tracklets degraded end-to-end performance, yielding relative drops of -1.0 to -2.5 mAP compared to single-frame local clustering. These preliminary results suggest that while this temporal strategy holds significant potential for geometric stabilisation, further work exploring how to mitigate noisy tracklet associations or how to design global optimisation algorithms robust to such noise remains an open challenge.

E.2 Anchor-Free Optimisation on Sparse Point Clouds

A core limitation of the pseudo-label generator is its reliance on categorical dimensional anchors to bound the PSO search space. To test if the pipeline could operate completely free of semantic biases, preliminary attempts were made to remove these anchors and perform purely unconstrained PSO on the local clusters. To initialise the swarm without a dimensional prior, we used Principal Component Analysis (PCA) on the raw point clusters to estimate the primary geometric axes and extents, following geometric methodologies explored in recent works [24].

However, this unconstrained approach failed primarily due to the inherent instability of spatial clustering heuristics and the non-uniformity of outdoor LiDAR point clouds. As established throughout this research, density-based clustering is highly prone to over- or under-segmentation.

Consequently, the raw point clusters fed into the PCA rarely represented the true, complete geometry of the object, yielding, therefore, poor results.

Furthermore, even on cleanly isolated objects, sparse LiDAR frequently yields hollow, non-uniform distributions (e.g., hitting only the L-shape of a vehicle's rear and side), which skews the computed PCA axes. Without categorical dimensional bounds to restrict the search space and correct these fundamentally flawed initialisations, the PSO algorithm failed to generalise to true object volumes.

Appendix F

Paper Inclusion Authorisation

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Co-Author Declaration and Authorisation

We, the undersigned, co-authors of the paper titled *Decoupling Geometry and Semantics for Robust Open-Vocabulary 3D Object Detection*, hereby declare and confirm that Pedro Rafael Angélico Madureira is the first and primary author of the aforementioned manuscript. We expressly authorise him to include and adapt text from this paper within his dissertation. We acknowledge that all implementations and evaluations presented in the dissertation were conducted independently and represent the exclusive work of Pedro Rafael Angélico Madureira.

By signing below, we agree to these terms and formally validate the independent nature of the work presented in his dissertation.

Assinado por: **Luís Filipe Pinto de Almeida Teixeira**
Num. de Identificação: 11308558
Data: 2026.09.18 10:28:38 +0100

Luís Filipe Teixeira

Assinado por: **José Diogo Xavier Monteiro**
Num. de Identificação: 14234515
Data: 2026.09.18 10:49:04 +0100

José Monteiro

Figure F.1: Copy of the co-authors' document authorising the inclusion and adaptation of the paper's content into this dissertation. The original document, digitally signed by all involved parties, can be requested by contacting the main author.