Group3D: MLLM-Driven Semantic Grouping for Open-Vocabulary 3D Object Detection

Youbin Kim[1], Jinho Park[1], Hogun Park[1], and Eunbyung Park[2]

[1] Department of Artificial Intelligence, Sungkyunkwan University
[2] Department of Artificial Intelligence, Yonsei University

Abstract. Open-vocabulary 3D object detection aims to localize and recognize objects beyond a fixed training taxonomy. In multi-view RGB settings, recent approaches often decouple geometry-based instance construction from semantic labeling, generating class-agnostic fragments and assigning open-vocabulary categories post hoc. While flexible, such decoupling leaves instance construction governed primarily by geometric consistency, without semantic constraints during merging. When geometric evidence is view-dependent and incomplete, this geometry-only merging can lead to irreversible association errors, including over-merging of distinct objects or fragmentation of a single instance. We propose Group3D, a multi-view open-vocabulary 3D detection framework that integrates semantic constraints directly into the instance construction process. Group3D maintains a scene-adaptive vocabulary derived from a multimodal large language model (MLLM) and organizes it into semantic compatibility groups that encode plausible cross-view category equivalence. These groups act as merge-time constraints: 3D fragments are associated only when they satisfy both semantic compatibility and geometric consistency. This semantically gated merging mitigates geometry-driven over-merging while absorbing multi-view category variability. Group3D supports both pose-known and pose-free settings, relying only on RGB observations. Experiments on ScanNet and ARKitScenes demonstrate that Group3D achieves state-of-the-art performance in multi-view open-vocabulary 3D detection, while exhibiting strong generalization in zero-shot scenarios. The project page is available at `https://ubin108.github.io/Group3D/`.

Keywords: 3D Object Detection · Open-Vocabulary · Multi-modal Large Language Model

1 Introduction

3D object detection aims to localize object instances in a scene while jointly estimating their 3D position, spatial extent, and semantic identity. Beyond pixel-/point-level scene interpretation, it provides structured, object-centric representations that serve as actionable abstractions of physical environments. Such representations are a core component of modern 3D perception, enabling explicit reasoning about object geometry and spatial relationships. As language

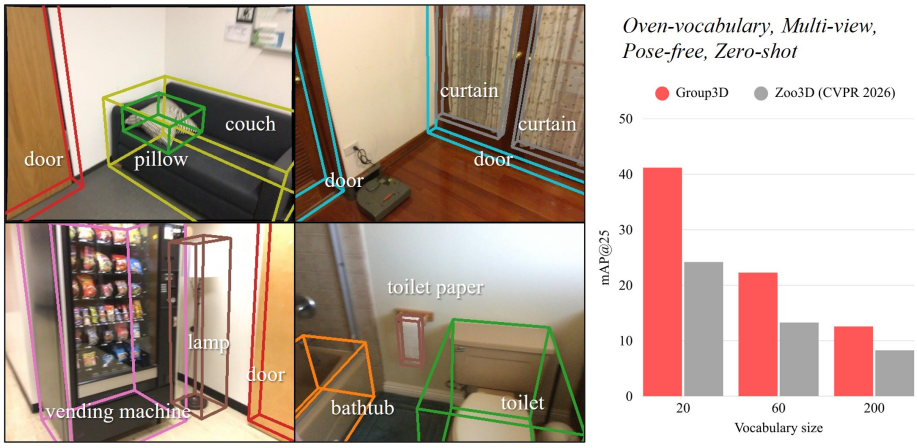

Fig. 1: Left: Predicted 3D bounding boxes projected onto the input RGB images. **Right**: Comparison with the baseline under the *multi-view, pose-free, zero-shot* setting across different vocabulary sizes, where Group3D consistently achieves higher mAP_{25}.

becomes increasingly intertwined with visual perception, grounding text-defined concepts to concrete 3D object instances further highlights the need for reliable instance-level 3D representations that support open-world perception.

Continuous advances in 3D geometric representation learning [6,16,23,27,30, 49,56] and instance-level localization strategies [25,29,38,53] have substantially improved accuracy and robustness of modern 3D object detectors. Yet most existing systems are still trained within a fixed label space defined by a predefined category taxonomy and dense 3D bounding-box annotations. Consequently, detectors remain tightly coupled to the training vocabulary, and extending recognition to new object types typically requires collecting and annotating additional 3D boxes—making scale-up costly and slow.

Open-vocabulary 3D object detection mitigates this limitation by relaxing the dependence on a fixed training taxonomy and enabling recognition beyond predefined class lists. In 2D, such capability has been enabled by large-scale vision–language alignment models [14, 32], which learn transferable semantics from image–text data. Extending this paradigm to 3D, existing approaches often transfer open-vocabulary signals from 2D models to generate pseudo 3D supervision for training 3D detectors. Although this reduces the need for manual 3D bounding box annotations, these pipelines generally assume access to explicit 3D geometry (e.g., point clouds) for proposal generation and localization. This assumption limits applicability in scenarios where acquiring dense 3D measurements is expensive or impractical.

As an alternative, multi-view image-based 3D detection leverages inexpensive and widely available RGB observations across views. Recent multi-view open-vocabulary 3D detection pipelines often construct 3D instances in a class-agnostic manner and incorporate semantic information only after instance formation or at the representation level. While such designs simplify open-vocabulary

labeling and maintain geometric robustness, they leave merging decisions governed primarily by geometric consistency. In multi-view RGB settings, geometric evidence is inherently view-dependent and often incomplete compared to ground-truth point clouds. As a result, geometry-driven merging under such ambiguity can fuse fragments that correspond to different semantic categories. Once boundaries are collapsed during instance construction, subsequent semantic reasoning may struggle to disentangle them reliably.

Building on this observation, we propose Group3D, a multi-view open vocabulary 3D object detection framework that integrates semantic and geometric cues during instance construction. Group3D operates on RGB observations of a single indoor scene and predicts a set of 3D object instances with open-vocabulary categories and 3D bounding boxes. Importantly, our approach is applicable in both pose-known and pose-free settings: when camera poses are available, Group3D directly leverages them for 3D lifting, while in the more challenging pose-free case it relies on reconstruction-based pose and depth estimates. Across both settings, the key objective is to prevent irreversible instance construction errors caused by incomplete or view-dependent geometry by enforcing semantic compatibility at merge time rather than only after instances are formed.

Group3D builds two scene-level memories to support open-vocabulary instance formation. First, it constructs a Scene Vocabulary Memory by querying a multimodal large language model (MLLM) across views, and aggregating them into a scene-adaptive vocabulary. Second, it constructs a 3D Fragment Memory by lifting category-aware 2D masks into 3D using multi-view geometry. This yields 3D fragments that preserve category hypotheses, confidence, and provenance, providing the atomic units for downstream instance construction.

Crucially, Group3D uses the MLLM to partition the scene vocabulary into semantic compatibility groups that capture plausible cross-view category variability. These groups induce a category-to-group mapping that gates fragment association. During instance formation, fragments are merged only when they satisfy both semantic compatibility and voxel-level geometric consistency. The resulting instances aggregate multi-view category evidence via confidence-weighted support statistics to select final open-vocabulary categories. As a result, Group3D achieves state-of-the-art performance in multi-view open-vocabulary 3D detection on both ScanNet [8] and ARKitScenes [2], while exhibiting strong zero-shot generalization. In summary, our contributions are summarized as follows:

- We propose Group3D, a multi-view open-vocabulary 3D detection framework that constructs instances by jointly leveraging semantic compatibility and geometric consistency, mitigating irreversible over-merging under geometric ambiguity.
- We introduce a novel MLLM-driven semantic grouping mechanism that exploits both open-vocabulary category prediction and language-induced compatibility priors to explicitly regulate 3D fragment association.
- We achieve strong open-vocabulary and zero-shot 3D detection performance using only multi-view RGB inputs, without requiring ground-truth depth or 3D supervision.

2 Related Works

2.1 3D Object Detection.

Point cloud-based detection. Early approaches processing point clouds to 3D object detection were confined to naively extending 2D detection paradigms into the 3D domain [30, 31, 38, 39, 52]. However, due to the sparsity of 3D data, this direct extension led to severe computational waste and significant bottlenecks in both detection speed and accuracy. To overcome this limitation, VoteNet [29] introduced a bottom-up architecture that integrated the Hough Voting into a deep learning framework. This has been established as the standard baseline for numerous closed-set 3D detectors [7, 43, 55]. Several methods [9, 10, 35, 37, 49, 50] further shifted its focus toward voxel-based paradigms. These approaches discretize the continuous 3D space into voxels, allowing for the direct application of efficient 3D convolutional operations.

Multi-view image-based detection. Multi-view image-based 3D detection constructs object representations from multiple RGB observations of a scene. These methods broadly encompass bird's-eye-view (BEV) projections [13, 18, 19, 21], DETR-based frameworks [5, 24, 41, 45]. Specifically, within the voxel-based paradigm, ImVoxelNet [36] constructs a 3D feature volume by directly lifting 2D image features into 3D voxel grids. Building upon this foundation, recent works [12, 20, 42, 47] have significantly advanced this approach. To further optimize the process, some methods [48, 54] explicitly predict and model the underlying scene geometry directly during the 2D-to-3D feature lifting phase. Despite these advances, most existing multi-view 3D detection frameworks operate under a closed-set setting, where detectors are trained to recognize a predefined set of object categories.

2.2 Open-Vocabulary 3D Object Detection.

Point cloud-based detection. A large body of work extends conventional 3D detectors to support open-vocabulary recognition using point cloud inputs. Early approaches adopt CLIP-style semantic transfer by aligning proposal features with text embeddings [32, 57]. Subsequent methods [3, 15, 26, 28, 46, 51] further improve detection by training open-vocabulary 3D detectors with pseudo supervision derived from 2D priors and cross-modal alignment. While these approaches significantly improve open-vocabulary recognition, they typically require training on target-domain data and rely primarily on geometry-driven instance association.

Multi-view image-based detection. Recent work has begun to extend multi-view image pipelines to open-vocabulary 3D detection. In these approaches, 2D predictions are lifted into 3D and aggregated across views to form object hypotheses. OpenM3D [11] proposes an open-vocabulary multi-view detection

framework trained with pseudo 3D boxes and CLIP-based semantic alignment without requiring human annotations. Zoo3D [17], in contrast, constructs 3D boxes by clustering lifted 2D masks and assigns semantic labels via vision-language similarity. However, these pipelines largely rely on geometric consistency for cross-view instance construction and incorporate semantic cues only after instances are formed. Geometry-first aggregation can lead to over-merging when observations are incomplete or geometrically ambiguous. Our method instead integrates semantic constraints directly into the instance construction process via MLLM-driven compatibility grouping, enabling more robust cross-view association.

3 Group3D

Problem Setup We address multi-view open-vocabulary 3D object detection from RGB observations. Given a set of RGB images $\mathcal{I} = \{I_n\}$ captured from a single scene, along with optional camera poses $\{T_n\}$, our goal is to predict a set of 3D object instances $\mathcal{O} = \{(\ell_k, s_k, b_k)\}_k$, where ℓ_k denotes the predicted open-vocabulary category, s_k is its confidence score, and b_k is an axis-aligned 3D bounding box.

3.1 Scene Memory Construction

Group3D constructs two scene-level memories: (i) *Scene Vocabulary Memory*, which aggregates object category hypotheses predicted across views into a compact scene-adaptive category set, and (ii) *3D Fragment Memory*, which stores all 3D fragments obtained by lifting category-aware 2D masks into the reconstructed 3D space.

Scene Vocabulary Memory. Given an input view I_n, we query an MLLM to obtain a set of object categories, $\mathcal{V}_n$. The predicted categories are normalized through canonicalization, including casing normalization and morphological standardization, e.g., *Trash_ can* $\rightarrow$ *trash can*. We then aggregate the normalized categories across views and remove duplicates to form a scene-level vocabulary, $\mathcal{V} = \bigcup_n \mathcal{V}_n$, referred to as the *Scene Vocabulary Memory*, which is subsequently used to induce semantic compatibility groups (Sec. 3.2).

3D Fragment Memory. We leverage a foundational segmentation model, SAM 3 [4] to obtain category-aware 2D masks. By querying each category $\ell_i \in \mathcal{V}$ in the scene vocabulary, we produce 2D masks $m_{n,i} \in \{0,1\}^{H \times W}$ for each input image I_n and each category ℓ_i, along with the confidence score $s_{n,i}$. Then, to lift 2D masks into 3D space, we obtain camera poses and depth maps using a reconstruction model applied to the input images. When ground-truth camera poses are available, we use them instead of the predicted poses. The resulting

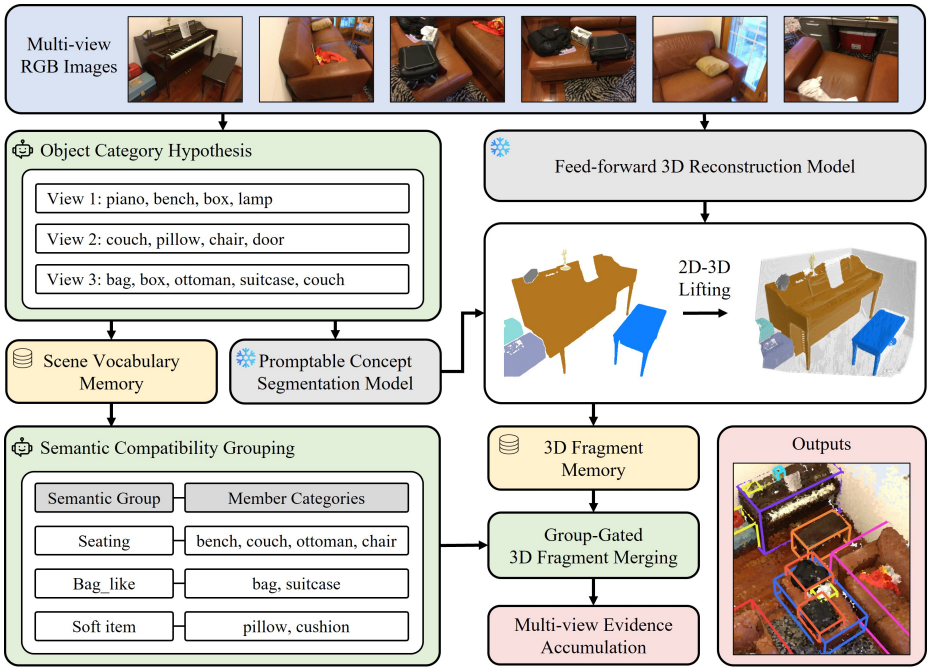

Fig. 2: The overview of Group3D. Given multi-view RGB images, an MLLM predicts object categories across views, which are aggregated into a *Scene Vocabulary Memory*. Category-aware masks are lifted into 3D to construct a *3D Fragment Memory*. The MLLM then organizes the vocabulary into semantic compatibility groups, which gate fragment merging together with geometric consistency to produce the final open-vocabulary 3D object instances. Finally, multi-view evidence is accumulated to determine the final open-vocabulary category and 3D bounding box for each object instance.

poses $\{T_n\}$ and depth maps $\{D_n\}$ define a shared world coordinate system for projecting 2D masks into 3D.

Each mask $m_{n,i}$ is lifted into 3D by back-projecting its pixel coordinates $\{(u,v) \mid m_{n,i}(u,v) = 1\}$ using the obtained depth and pose, where $(\cdot,\cdot)$ denotes an indexing operator. Let K_n denote the camera intrinsic matrix and $T_n = [R_n \mid t_n]$ the camera pose mapping world coordinates to the camera frame. We lift each mask $m_{n,i}$ into 3D via back-projection,

$$p(u,v) = R_n^\top \left(D_n(u,v)\, K_n^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} - t_n \right), \tag{1}$$

and define the corresponding point clouds fragment $F_{n,i}$, and *3D Fragment Memory* $\mathcal{F}$ can be defined as follows,

$$\mathcal{F} := \{(F_{n,i}, \ell_i, s_{n,i})\}_{n,i}, \quad F_{n,i} = \{p(u,v) \in \mathbb{R}^3 \mid m_{n,i}(u,v) = 1, \forall u, v\}. \tag{2}$$

To mitigate reconstruction noise, we apply reliability filtering and suppress extreme depth outliers within each mask region, and each fragment stores its 3D point clouds $F_{n,i}$, category hypothesis ℓ_i, and confidence score $s_{n,i}$.

Regarding the confidence score, we defined it as the product of a query-level score and a global presence score:

$$s_{n,i} = s_{n,i}^{\text{query}} \cdot s_n^{\text{pres}}, \tag{3}$$

where s_n^{pres} estimates whether the prompted category is present in I_n, and $s_{n,i}^{\text{query}}$ measures the match of the corresponding between the prompted category and the mask region.

3.2 Semantic Compatibility Grouping

Open-vocabulary predictions across views can be inconsistent due to taxonomy noise, where the same physical object may receive different but semantically related categories across frames. To transform this variability into a structured prior for instance construction, we query the MLLM to partition the scene vocabulary $\mathcal{V}$ into semantic compatibility groups,

$$\mathcal{G} = \{G_g\}_{g=1}^{G}, \quad G_g \subseteq \mathcal{V}. \tag{4}$$

The MLLM is prompted to group categories that could plausibly refer to the same physical object under taxonomy noise (e.g., *chair–sofa*, *desk–table*), while avoiding merges that are structurally inconsistent. In particular, categories corresponding to structural attachments (e.g., *wall–window* or *wall–door*), supporting structures (e.g., *floor–wall*), or part–whole relationships (e.g., *table–cup*) are explicitly excluded from the same group.

As a result, the induced grouping captures semantic substitutability rather than spatial adjacency or co-occurrence. These groups define which category labels are considered compatible and therefore allowed to merge across views. Candidate merges are subsequently verified using geometric consistency during instance merging.

3.3 Group-Gated 3D Fragment Merging

We construct global 3D instances by merging fragments in $\mathcal{F}$ under a semantic compatibility constraint combined with geometric consistency. The defining characteristic of Group3D is that fragment association is explicitly gated by semantic compatibility groups introduced in Sec. 3.2, rather than relying on geometry alone. Let $g(\cdot)$ denote the mapping a category (or a set of categories) to semantic compatibility group, then two point cloud fragments $F_{n,i}$ and $F_{m,j}$ are merge-eligible only if they satisfy the following condition $g(\ell_i) = g(\ell_j)$, i.e., both categories are in the same group, ensuring that only semantically compatible fragments can be associated.

Algorithm 1: Group-Gated 3D Fragment Merging

Input: *3D Fragments Memory* $\mathcal{F}$
Output: Merged clusters $\mathcal{C}$ (3D instances)
Sort fragments by descending spatial extent;
$\mathcal{C} \leftarrow \emptyset$;
foreach $(F_{n,i}, \ell_i, s_{n,i}) \in \mathcal{F}$ **do**
  merged $\leftarrow$ False;
  foreach $(C_F, C_\ell) \in \mathcal{C}$ **do**
    if $g(\ell_i) = g(C_\ell) \ \wedge \ \texttt{Overlap}(F_{n,i}, C_F)$ **then**
      $C_F \leftarrow C_F \cup \{F_{n,i}\}$; `// Merge point clouds fragments`
      $C_\ell \leftarrow C_\ell \cup \{\ell_i\}$; `// Keep a set of open-vocab categories`
      merged $\leftarrow$ True;
      break;
  if merged = False **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{(F_{n,i}, \{\ell_i\})\}$; `// Add a new fragment`
return $\mathcal{C}$;

Geometric consistency is then verified using voxel overlap. Each fragment is represented by its voxel set $\texttt{vox}(\cdot)$, and overlap is measured using Intersection over Union (IoU) together with a containment ratio:

$$\mathrm{IoU}_{\texttt{vox}}(A, B) = \frac{|\texttt{vox}(A) \cap \texttt{vox}(B)|}{|\texttt{vox}(A) \cup \texttt{vox}(B)|}, \qquad \mathrm{Cont}_{\texttt{vox}}(B \to A) = \frac{|\texttt{vox}(A) \cap \texttt{vox}(B)|}{|\texttt{vox}(B)|}. \tag{5}$$

IoU alone may underestimate geometric agreement when fragments differ substantially in spatial extent. If fragment B is significantly smaller than fragment A, IoU can remain low even when B overlaps heavily with A or is almost entirely contained within it. In such cases, the union term is dominated by the larger fragment, diluting the overlap score. The containment ratio explicitly measures how much of the smaller fragment is supported by the larger one, thereby capturing this asymmetric inclusion. We define $\texttt{Overlap}(A, B)$ as a boolean predicate based on these measures as follows,

$$\mathrm{Overlap}(A, B) = (\mathrm{IoU}_{\mathrm{vox}}(A, B) \geq \tau_{\mathrm{iou}}) \ \vee \ (\mathrm{Cont}_{\mathrm{vox}}(B \to A) \geq \tau_{\mathrm{cont}}). \tag{6}$$

Combining these conditions, the fragment merging is performed under the conjunction of semantic compatibility, geometric overlap, and cross-view consistency. Algorithm 1 summarizes the resulting group-gated merging procedure, which produces final 3D instance clusters $\mathcal{C}$.

3.4 Multi-view Evidence Accumulation

After group-gated merging (Algorithm 1), each 3D instance (C_F, C_ℓ) contains the merged point cloud fragments C_F and the set of associated category labels

C_ℓ. To determine the final label, we aggregate the candidate categories together with their confidence scores. Slightly abusing notation, let $\ell \in C_\ell$ denote a category label associated with the 3D instance. We compute its mean confidence score $\bar{s}(\ell)$ by averaging the confidence scores of all fragments associated with the category ℓ. Since each 3D instance is formed by merging fragments originating from multiple input views, the same category may be associated with multiple fragments, each carrying a different confidence score. The instance-level category score is then defined as,

$$s(\ell) = \bar{s}(\ell) \cdot w(N(\ell)) \,, \tag{7}$$

where $N(\ell)$ denotes the number of fragments associated with category ℓ during the merging stage, and $w(x) = 1 - \exp\left(-\frac{x}{\tau}\right)$ is a monotonically increasing function that rewards repeated cross-view support while preventing disproportionate dominance by categories with many fragments. The final instance label is selected as $\arg\max_{\ell \in C_\ell} s(\ell)$, with the corresponding score $s(\ell)$. The 3D bounding box is computed by taking the minimum and maximum coordinates of C_F along each axis.

4 Experiments

4.1 Datasets

Evaluation is conducted on two multi-view indoor 3D perception benchmarks, ScanNetV2 [8] and ARKitScenes [2], and results are reported on the official validation splits. Since the proposed pipeline is training-free with respect to 3D supervision, these benchmarks are used solely for evaluation. Following standard 3D object detection protocols, mean average precision (mAP) is reported at 3D IoU thresholds of 0.25 and 0.50.

ScanNet. ScanNetV2 [8] is a standard indoor RGB-D benchmark that provides reconstructed scenes with multi-view RGB sequences, aligned camera trajectories, and 3D instance annotations. The official split contains 1,201 training scenes and 312 validation scenes. To characterize open-vocabulary generalization across vocabulary scales, three established settings are considered, denoted as ScanNet20, ScanNet60, and ScanNet200: (i) a 20-category setting following [26]; (ii) a 60-category setting following [3, 46], where categories are defined by training frequency, treating the top-10 most frequent categories as seen and 50 additional categories as novel; and (iii) a 200-category setting following [34], which expands the label space to 200 fine-grained categories with a pronounced long-tail distribution. The ScanNet60 setting is commonly used with supervised training on the seen categories; comparisons therefore include both supervised methods trained on the seen set and zero-shot methods that use no category-specific 3D supervision.

Table 1: Quantitative results on the ScanNet benchmark under two category settings. Methods are grouped by input modality, including point cloud-based methods and multi-view image-based methods. For multi-view methods, results are further reported with and without ground-truth camera poses. † denotes methods that use 3D bounding boxes during training. Zoo3D$_0$ and Zoo3D$_1$ denote the zero-shot and self-supervised variants of Zoo3D, respectively.

Method	Pose-free	Zero-shot	ScanNet20		ScanNet60	
			mAP$_{25}$	**mAP$_{50}$**	**mAP$_{25}$**	**mAP$_{50}$**
Point cloud-based						
Det-PointCLIPv2† [57]	–	✗	–	–	0.2	–
3D-CLIP† [32]	–	✗	–	–	4.0	–
OV-3DET [26]	–	✗	18.0	–	–	–
CoDa† [3]	–	✗	19.3	–	9.0	–
INHA† [15]	–	✗	–	–	10.7	–
GLIS [28]	–	✗	20.8	–	–	–
ImOV3D [51]	–	✗	21.5	–	–	–
OV-Uni3DETR† [46]	–	✗	25.3	–	19.4	–
Zoo3D$_0$ [17]	–	✓	34.7	23.9	27.1	18.7
Zoo3D$_1$ [17]	–	✗	**37.2**	**26.3**	**32.0**	**20.8**
Multi-view image-based						
OV-Uni3DETR† [46]	✗	✗	–	–	11.2	–
OpenM3D [11]	✗	✗	19.8	7.3	–	–
Zoo3D$_0$ [17]	✗	✓	30.5	17.3	22.0	10.4
Zoo3D$_1$ [17]	✗	✗	32.8	15.5	23.9	10.8
Group3D (Ours)	✗	✓	**51.1**	**27.4**	**29.1**	**13.9**
Zoo3D$_0$ [17]	✓	✓	24.2	8.8	13.3	4.1
Zoo3D$_1$ [17]	✓	✗	27.9	10.4	15.3	5.6
Group3D (Ours)	✓	✓	**41.2**	**18.5**	**22.3**	**8.5**

ARKitScenes. ARKitScenes [2] provides real-world indoor multi-view RGB-D sequences with reconstructed scene geometry and 3D object annotations for 17 object categories. The official split contains 4,493 training scans and 549 validation scans.

4.2 Implementation Details

Experimental settings. For each scene, we uniformly sample 128 frames and resize all frames to 378×504 for reconstruction, following the input resolution setting of the reconstruction backbone. We extract the K category hypotheses per view for scene vocabulary construction and set $K = 5$ in all experiments. We use GPT-5.1 as the MLLM for category proposal and semantic grouping. During group-gated fragment merging, we voxelize fragments with a fixed voxel size of $5cm$ to compute voxel overlap and containment. All experiments are conducted on a single NVIDIA A6000 GPU. Additional implementation details are provided in the supplementary material.

Zero-shot setting. All results are obtained in a zero-shot manner without using category-specific 3D supervision from the evaluated benchmarks. To avoid

Table 2: Per-class AP_{25} comparison on ScanNet20. 'pc' indicates the methods leverage the ground-truth point clouds, 'pi' denotes the use of posed images, and 'ui' means the use of unposed images.

	Inputs	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink	
OV-3DET	pc + pi	57.3	42.3	27.1	31.5	8.2	14.2	3.0	5.6	23.0	31.6	
CoDA	pc	68.1	44.1	28.7	44.6	3.4	20.2	5.3	0.1	28.0	45.3	
OV-Uni3DETR	pc	86.1	50.5	28.1	31.5	18.2	24.0	6.6	12.2	29.6	54.6	
Zoo3D$_1$	pc + pi	78.4	54.4	**74.4**	65.5	33.6	19.1	14.1	32.3	**46.1**	27.3	
Group3D	pi	**91.3**	**80.5**	61.1	**78.0**	**55.4**	**56.2**	**24.7**	**41.3**	43.7	**62.3**	
(Ours)	ui	89.6	80.2	37.8	73.4	50.5	43.1	16.6	31.7	20.2	39.3	

		bathtub	refrigerator	desk	nightstand	counter	door	curtain	box	lamp	bag	Mean
OV-3DET		56.3	11.0	19.7	0.8	0.3	9.6	10.5	3.8	2.1	2.7	18.0
CoDA		50.5	6.6	12.4	15.2	0.7	8.0	0.0	2.9	0.5	2.0	19.3
OV-Uni3DETR		63.7	14.4	30.5	2.9	1.0	1.0	19.9	12.7	5.6	13.5	25.3
Zoo3D$_1$		64.6	57.5	10.7	58.8	0.2	27.4	8.0	20.0	**43.3**	9.1	37.2
Group3D		**85.8**	**63.3**	**65.1**	**70.4**	**2.0**	**40.8**	**24.1**	**22.6**	35.7	**18.3**	**51.1**
(Ours)		82.4	52.2	60.1	57.8	0.9	24.0	20.9	15.8	13.7	14.4	41.2

Table 3: Quantitative results on ScanNet200 [34] and ARKitScenes [2] under the multi-view setting.

Method	**Pose-free**	**Zero-shot**	**ScanNet200**		**ARKitScenes**	
			mAP$_{25}$	**mAP$_{50}$**	**mAP$_{25}$**	**mAP$_{50}$**
OpenM3D [11]	✗	✗	4.2	–	–	–
Zoo3D$_0$ [17]	✗	✓	14.3	6.2	–	–
Zoo3D$_1$ [17]	✗	✗	16.5	6.3	–	–
Ours	✗	✓	**17.9**	**8.7**	**20.5**	**5.9**
Zoo3D$_0$ [17]	✓	✓	8.3	2.9	13.0	2.6
Zoo3D$_1$ [17]	✓	✗	10.7	3.8	16.1	3.5
Ours	✓	✓	**12.6**	**5.7**	**18.4**	**4.5**

dataset-specific training leakage, 3D reconstruction backbones are selected such that they are not trained on the target benchmark. Accordingly, Depth Anything 3 [22] is used for ScanNetV2, and VGGT [44] is used for ARKitScenes. This ensures that the proposed pipeline does not rely on dataset-specific supervision from the evaluation benchmarks.

Reconstruction-based geometry and alignment. Both pose-known and pose-free settings rely on RGB-only reconstruction for 3D lifting; in the pose-known setting, the provided camera poses are used in place of estimated poses. Following Zoo3D [17], we align the reconstructed geometry to the benchmark coordinate system by matching the first predicted pose to the first ground-truth pose and calibrating the global scale using the first-frame depth.

4.3 Results and Comparisons

We compare Group3D with prior open-vocabulary 3D detection approaches under two regimes, *pose-known* and *pose-free*. We focus exclusively on the multi-

view RGB setting and report comparisons to both point cloud-based open-vocabulary detectors and multi-view image-based pipelines (Tab. 1).

On ScanNet20, Group3D establishes a clear new state-of-the-art among multi-view methods. It also surpasses representative approaches that rely on ground-truth point clouds, which highlights that semantically constrained instance construction can compensate for the absence of explicit 3D measurements and even outperform stronger-input baselines. On ScanNet60, Group3D remains competitive and improves over existing multi-view RGB pipelines, suggesting that semantic compatibility grouping continues to provide useful constraints as the vocabulary expands. We additionally evaluate on ScanNet200, a more challenging long-tail setting with a substantially larger and finer-grained vocabulary (Tab. 3). Group3D remains effective under this expanded category space, indicating that MLLM-driven grouping scales beyond a compact taxonomy and supports open-vocabulary recognition in the presence of long-tail categories, as shown in Fig. 4. In the pose-free regime, where reconstruction noise makes geometry-only association particularly brittle, Group3D preserves strong performance and demonstrates robustness when geometric evidence is incomplete or uncertain. These results highlight the benefit of incorporating semantic compatibility during instance construction, particularly in scenarios where geometric cues alone are insufficient to reliably associate fragments across views. Notably, these improvements are achieved without relying on explicit 3D supervision or ground-truth depth measurements. This suggests that combining multi-view geometric cues with language-driven semantic constraints can provide a viable alternative to conventional geometry-driven pipelines.

Table 2 reports per-class results on ScanNet20 and shows that Group3D achieves consistent improvements across a broad set of categories, while Fig. 3 provides visualizations of the predicted open-vocabulary 3D detections. Results on ARKitScenes (Tab. 3) further demonstrate that Group3D generalizes to a different dataset with distinct capture conditions and scene statistics, suggesting that the semantic compatibility prior transfers across domains and supports open-vocabulary 3D detection in the wild.

Table 4: Ablation study on ScanNet20 with different components. Depth Anything 3 is trained on external datasets, while VGGT is pretrained on ScanNet.

Component	**Method**	**mAP**$_{25}$	**mAP**$_{50}$
Reconstruction	Depth Anything 3 [22]	41.2	18.5
	VGGT [44]	40.0	18.7
MLLM	GPT 5.1 [40]	41.2	18.5
	Qwen3-VL-8B [1]	38.5	16.9
Segmentation	SAM 3 [4]	41.2	18.5
	Grounded SAM 2 [33]	39.7	17.6

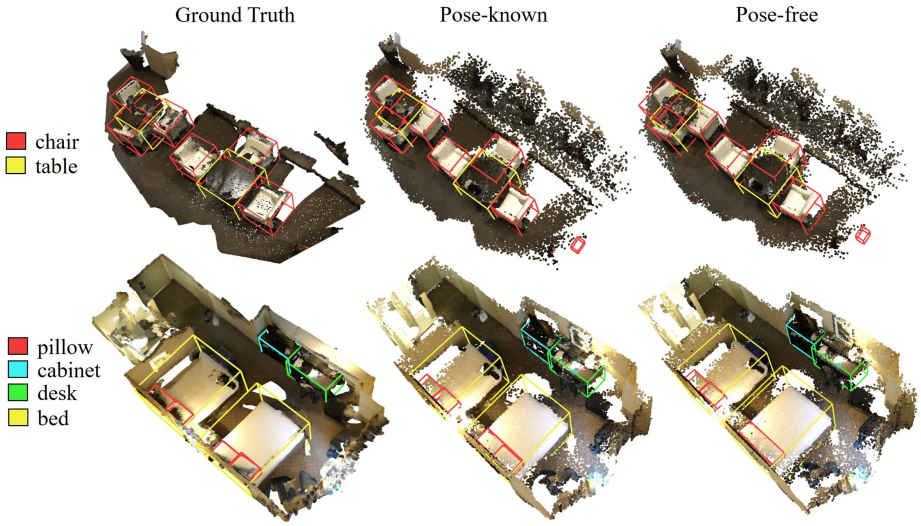

Fig. 3: Qualitative results on ScanNet20 [26] under *pose-known* and *pose-free* settings.

Table 5: Ablation on ScanNet20 with proposed components for fragment merging.

Component	$\mathbf{mAP}_{25}$	$\mathbf{mAP}_{50}$
Geometry-only	28.2	9.9
+ Scene Vocabulary Memory	35.9	14.8
+ Semantic Compatibility Grouping	41.2	18.5

Table 6: Ablation on ScanNet20 with different K of object category hypotheses.

K	$\mathbf{mAP}_{25}$	$\mathbf{mAP}_{50}$
$K = 5$	41.2	18.5
$K = 10$	41.2	18.8

4.4 Ablation Study

We analyze the key components of Group3D on ScanNet20. As shown in Tab. 6, varying the number of category hypotheses per view has limited impact on performance: using $K = 5$ or $K = 10$ yields comparable results. We therefore adopt $K = 5$ for improved efficiency without sacrificing accuracy.

Tab. 4 compares different reconstruction backbones, MLLMs, and segmentation models. Replacing the MLLM with a smaller 8B-scale model leads to a moderate performance drop, yet the overall pipeline remains effective, indicating that the proposed semantic grouping mechanism is not tightly coupled to a specific large-scale language model. For reconstruction, VGGT achieves competitive performance but is trained on ScanNet, whereas Depth Anything 3 maintains strong results in a strictly zero-shot setting. Replacing SAM 3 with Grounded SAM 2 slightly reduces performance due to differences in grounding and confidence formulation, while preserving the overall trend.

Tab. 5 underscores the importance of each proposed component during fragment merging. Removing both components and merging purely by geometry leads to clear degradation due to geometry-driven over-merging. Adding Scene Vocabulary Memory enforces a same-category constraint during merging, miti-

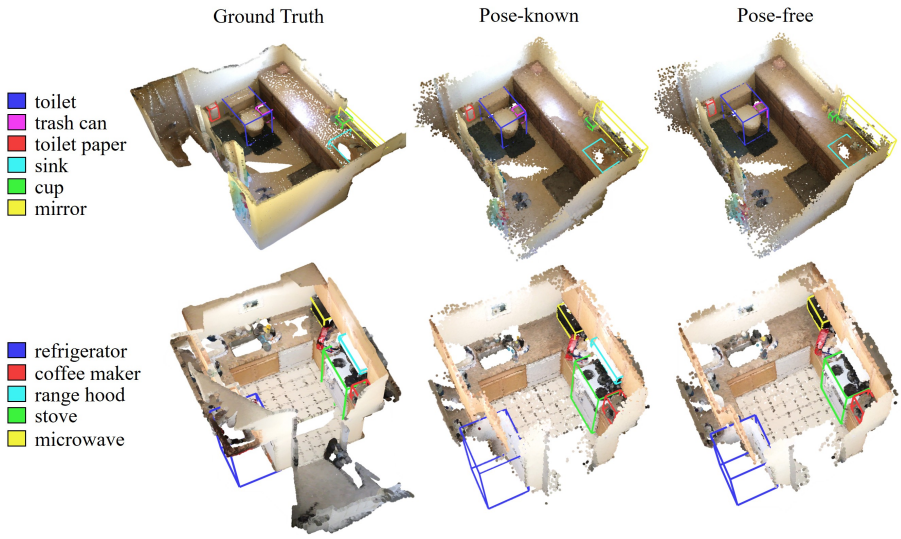

Fig. 4: Qualitative results on ScanNet200 [34] under *pose-known* and *pose-free* settings.

gating over-merging errors but remaining sensitive to cross-view label variability. Further adding Semantic Compatibility Grouping addresses this limitation by allowing semantically consistent labels to merge while preventing structurally incompatible associations, achieving the best performance.

5 Conclusion

In this work, we proposed Group3D, a multi-view open-vocabulary 3D object detection framework that incorporates semantic constraints directly into the instance construction process. By organizing scene-adaptive category hypotheses into semantic compatibility groups and enforcing merge-time semantic gating, Group3D mitigates geometry-driven over-merging under incomplete and view-dependent multi-view evidence while remaining robust to cross-view category variability. Overall, the results suggest that injecting semantic compatibility into fragment merging leads to more reliable open-vocabulary 3D instance construction using only multi-view RGB inputs.

More broadly, our findings suggest that integrating language-driven semantic priors into the instance construction process can complement geometric reasoning in multi-view 3D perception. Such integration may provide a scalable pathway toward open-world 3D scene understanding without relying on dense 3D supervision or explicit geometry sensors. We hope this perspective encourages further research on language-guided 3D perception. Future work may explore extending the framework to support richer language descriptions and more complex scene-level reasoning across objects, further strengthening the integration between language understanding and multi-view 3D perception.

Acknowledgements

This research was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT2401-01, and by the Ministry of Science and ICT (MSIT) of Korea, including the National Research Foundation (NRF) grant (RS-2024-00337548), the Advanced GPU Utilization Support Program, and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants under the following projects: AI Research Hub Project (No. RS-2024-00457882), Development of a human foundation model for human-centric universal artificial intelligence and training of personnel (No. RS-2025-25441838), and Development of Compression, Reconstruction, and Rendering Technologies for Free-viewpoint Media (No. RS-2026-25517417).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021)
3. Cao, Y., Yihan, Z., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. Advances in Neural Information Processing Systems **36**, 71862–71873 (2023)
4. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21674–21683 (2023)
7. Cheng, B., Sheng, L., Shi, S., Yang, M., Xu, D.: Back-tracing representative points for voting-based 3d object detection in point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8963–8972 (2021)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
9. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel r-cnn: Towards high performance voxel-based 3d object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1201–1209 (2021)
10. Gwak, J., Choy, C., Savarese, S.: Generative sparse detection networks for 3d single-shot object detection. In: European conference on computer vision. pp. 297–313. Springer (2020)

11. Hsu, P.H., Zhang, K., Wang, F.E., Tu, T., Li, M.F., Liu, Y.L., Chen, A.Y., Sun, M., Kuo, C.H.: Openm3d: Open vocabulary multi-view indoor 3d object detection without human annotations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8688–8698 (2025)
12. Huang, C., Hou, Y., Ye, W., Huang, D., Huang, X., Lin, B., Cai, D.: Nerf-det++: Incorporating semantic cues and perspective-aware depth supervision for indoor multi-view 3d detection. IEEE Transactions on Image Processing (2025)
13. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
15. Jiao, P., Zhao, N., Chen, J., Jiang, Y.G.: Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image. In: European Conference on Computer Vision. pp. 376–392. Springer (2024)
16. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
17. Lemeshko, A., Gabdullin, B., Drozdov, N., Konushin, A., Rukhovich, D., Kolodiazhnyi, M.: Zoo3d: Zero-shot 3d object detection at scene level. arXiv preprint arXiv:2511.20253 (2025)
18. Li, H., Zhang, H., Zeng, Z., Liu, S., Li, F., Ren, T., Zhang, L.: Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6684–6693 (2023)
19. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 1477–1485 (2023)
20. Li, Z., Yu, H., Ding, Y., Qiao, J., Azam, B., Akhtar, N.: Go-n3rdet: Geometry optimized nerf-enhanced 3d object detector. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27211–27221 (2025)
21. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 2020–2036 (2024)
22. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
23. Liu, S., Cui, M., Li, B., Liang, Q., Hong, T., Huang, K., Shan, Y.: Fshnet: Fully sparse hybrid network for 3d object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8900–8909 (2025)
24. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: European conference on computer vision. pp. 531–548. Springer (2022)
25. Liu, Z., Hou, J., Ye, X., Wang, T., Wang, J., Bai, X.: Seed: A simple and effective 3d detr in point clouds. In: European Conference on Computer Vision. pp. 110–126. Springer (2024)

26. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1190–1199 (2023)
27. Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., Xu, C.: Voxel transformer for 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3164–3173 (2021)
28. Peng, X., Bai, Y., Gao, C., Yang, L., Xia, F., Mu, B., Wang, X., Liu, S.: Global-local collaborative inference with llm for lidar-based open-vocabulary detection. In: European Conference on Computer Vision. pp. 367–384. Springer (2024)
29. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019)
30. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
31. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
33. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
34. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: European conference on computer vision. pp. 125–141. Springer (2022)
35. Rukhovich, D., Vorontsova, A., Konushin, A.: Fcaf3d: Fully convolutional anchor-free 3d object detection. In: European Conference on Computer Vision. pp. 477–493. Springer (2022)
36. Rukhovich, D., Vorontsova, A., Konushin, A.: Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2397–2406 (2022)
37. Rukhovich, D., Vorontsova, A., Konushin, A.: Tr3d: Towards real-time indoor 3d object detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 281–285. IEEE (2023)
38. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019)
39. Shi, W., Rajkumar, R.: Point-gnn: Graph neural network for 3d object detection in a point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1711–1719 (2020)
40. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
41. Tseng, C.Y., Chen, Y.R., Lee, H.Y., Wu, T.H., Chen, W.C., Hsu, W.H.: Crossdtr: Cross-view and depth-guided transformers for 3d object detection. In: 2023 IEEE

International Conference on Robotics and Automation (ICRA). pp. 4850–4857. IEEE (2023)
42. Tu, T., Chuang, S.P., Liu, Y.L., Sun, C., Zhang, K., Roy, D., Kuo, C.H., Sun, M.: Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6996–7007 (2023)
43. Wang, H., Shi, S., Yang, Z., Fang, R., Qian, Q., Li, H., Schiele, B., Wang, L.: Rbgnet: Ray-based grouping for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1110–1119 (2022)
44. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
45. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on robot learning. pp. 180–191. PMLR (2022)
46. Wang, Z., Li, Y., Liu, T., Zhao, H., Wang, S.: Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation. In: European Conference on Computer Vision. pp. 73–89. Springer (2024)
47. Xu, C., Wu, B., Hou, J., Tsai, S., Li, R., Wang, J., Zhan, W., He, Z., Vajda, P., Keutzer, K., et al.: Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 23320–23330 (2023)
48. Xu, Y., Li, C., Lee, G.H.: Mvsdet: Multi-view indoor 3d object detection via efficient plane sweeps. Advances in Neural Information Processing Systems **37**, 132824–132842 (2024)
49. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
50. Yang, B., Luo, W., Urtasun, R.: Pixor: Real-time 3d object detection from point clouds. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7652–7660 (2018)
51. Yang, T., Ju, Y., Yi, L.: Imov3d: Learning open vocabulary point clouds 3d object detection from only 2d images. Advances in Neural Information Processing Systems **37**, 141261–141291 (2024)
52. Yang, Z., Sun, Y., Liu, S., Shen, X., Jia, J.: Std: Sparse-to-dense 3d object detector for point cloud. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1951–1960 (2019)
53. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
54. Zhang, R., Yu, Z., Cao, S.Y., Zhu, L., Zhang, G., Bai, X., Shen, H.L.: Boosting multi-view indoor 3d object detection via adaptive 3d volume construction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5980–5989 (2025)
55. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3dnet: 3d object detection using hybrid geometric primitives. In: European conference on computer vision. pp. 311–329. Springer (2020)
56. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)

57. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2639–2650 (2023)